Neural Networks 155 (2022) 224–241

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

Deep learning-based image deconstruction method with maintained saliency

Keisuke Fujimoto [a], Kojiro Hayashi [a], Risa Katayama [a], Sehyung Lee [a], Zhen Liang [c,a], Wako Yoshida [a,d], Shin Ishii [a,b,*]

[a] Graduate School of Informatics, Kyoto University, Kyoto 606-8501, Japan
[b] ATR Neural Information Analysis Laboratories, Kyoto 619-0288, Japan
[c] School of Biomedical Engineering, Health Science Center, Shenzhen University, Shenzhen, People's Republic of China
[d] Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford, United Kingdom

ARTICLE INFO

Article history:
Received 23 November 2020
Received in revised form 30 June 2022
Accepted 12 August 2022
Available online 23 August 2022

Keywords:
Attention
Image transformation
Saliency map
Deep learning
Variational autoencoder
Functional magnetic resonance imaging

ABSTRACT

Visual properties that primarily attract bottom-up attention are collectively referred to as saliency. In this study, to understand the neural activity involved in top-down and bottom-up visual attention, we aim to prepare pairs of natural and unnatural images with common saliency. For this purpose, we propose an image transformation method based on deep neural networks that can generate new images while maintaining the consistent feature map, in particular the saliency map. This is an ill-posed problem because the transformation from an image to its corresponding feature map could be many-to-one, and in our particular case, the various images would share the same saliency map. Although stochastic image generation has the potential to solve such ill-posed problems, the most existing methods focus on adding diversity of the overall style/touch information while maintaining the naturalness of the generated images. To this end, we developed a new image transformation method that incorporates higher-dimensional latent variables so that the generated images appear unnatural with less context information but retain a high diversity of local image structures. Although such high-dimensional latent spaces are prone to collapse, we proposed a new regularization based on Kullback–Leibler divergence to avoid collapsing the latent distribution. We also conducted human experiments using our newly prepared natural and corresponding unnatural images to measure overt eye movements and functional magnetic resonance imaging, and found that those images induced distinctive neural activities related to top-down and bottom-up attentional processing.

1. Introduction

One of the remaining enigmas in the mammalian visual system is the mechanisms in dynamic control of its processing resources; the animal visual system is equipped with an efficient system to extract useful information from a tremendous amount of visual information input to the retina, that is, visual attention. Elucidating visual attention mechanisms is also important for developments of human harmonic systems; for example, the detection of driver's awareness/unawareness of pedestrians is serious for assuring the security of semi-automatic mobile systems (Dollar, Wojek, Schiele, & Perona, 2011), and general object detection with head-mounted cameras is an important issue for enlarging the applicability of agile robots to a variety of industrial scenes (Jiang et al., 2021; Weng et al., 2021). For awake animals, there are presumably two streams of attentional processing. One is bottom-up processing, which is assumed to process the characteristic portions of images or image-series with high priority. Saliency maps have been developed for comprehensively representing the image regions to attract mostly bottom-up attention as heat maps (Harel, Koch, & Perona, 2006; Itti, Koch, & Niebur, 1998). The other is top-down processing, which is assumed to process the contexts and scenes by checking them in light of prior knowledge and past experiences of individuals. There is a series of studies that discussed the relationship between the saliency map and voluntary eye movements (Itti, 2005, 2006; Veale, Hafed, & Yoshida, 2017; Yoshdia et al., 2010). Since the two kinds of processing above are mixed in the awake animal's visual processing, however, it has been a long-term challenge to dissociate them and then to examine them distinctively in research on mammalian visual processing.

Keeping the above background in mind, the purpose of this study is to develop an image transformation method for preparing a set of image pairs that have the same visual saliency map

* Correspondence to: 36-1 Yoshida-Honmachi, Sakyo-ku, Kyoto 606-8501, Japan.
E-mail address: ishii@i.kyoto-u.ac.jp (S. Ishii).

https://doi.org/10.1016/j.neunet.2022.08.015

but are natural on one side and unnatural images with minor context information on the other side. If the saliency map well represents the bottom-up attention, the generated and unnatural image would exhibit comparable or even better consistency with the eye movements in an awake and overt environment, but have less association with the top-down context. That is, we assume the top-down attention would be reduced to a large extent when we look at unnatural images with destroyed context information. The validity of this assumption was examined by behavioral experiments with human subjects. After checking the validity, we performed human experiments with functional magnetic resonance imaging (fMRI); the comparison between fMRI-based brain activities when looking at natural and unnatural images provides us with an insight into neural bases involved in top-down and bottom-up attention, respectively.

Although some prior works proposed to estimate the saliency map of a given natural image, based on deep learning techniques (Kruthiventi, Ayush, & Babu, 2017; Pan et al., 2017), there has been no study to propose an inverse transformation from a saliency map to natural/unnatural images. Since the transformation from an image to its saliency map is many-to-one, due to the lower dimensionality of the saliency map space, image generation constrained on a specific saliency map is a typical ill-posed problem. This observation would make the generation of multiple images that share the same saliency map seemingly easy, but it is not so in practice, because generated images are prone to lose their diversity due to mode collapse in general generative processes, or too complicated ones due to the high-dimensional artifacts usually introduced during the inverse process. Such generated images would not work for our purpose; we expect the image transformation transforms from a natural image to an unnatural one such that local structures therein are destroyed to enable human observers not to understand the context of the original images. To obtain such pairs, the diversity in terms of local structures of images is of our focus.

To achieve diversity in the image generation, existing studies introduced perturbations to the network, with stochastic masking such as Dropouts (Isola, Zhu, Zhou, & Efros, 2017) and stochastic sampling at the level of intermediate representations (Kingma & Welling, 2013). Among those, BicycleGAN (Zhu et al., 2017) successfully generated pairs of images with different styles, using stochastic sampling based on a combination of variational autoencoder (VAE) (Kingma & Welling, 2013) and generative adversarial network (GAN) (Goodfellow et al., 2014). The same authors utilized a technique of latent regressor (LR) to realize a diverse transformation between the image pair. Although this method is suitable for image generation with different styles, it could not be used for our purpose, because its skip connections (Ronneberger, Fischer, & Brox, 2015) worked for making local structures consistent between the pair of images by preserving the spatially local image features. This is an advantage when transforming styles or touches, which are mostly global features of images, but is a disadvantage when generating unnatural images in which local structures such as local shapes or colors must be collapsed to destroy context information.

In the present study, we developed a new image transformation method based on deep neural networks (DNNs), that transforms a given natural image to an unnatural and deconstructed image, while their saliency maps are kept very similar. Considering possible applications to examining the human visual attentional system, we needed to emerge diversity in local structures of generated images. To this end, we proposed a new Kullback–Leibler (KL) divergence-based regularization to prevent the latent distribution of relatively large degrees of freedom (DOFs) from being collapsed. In the existing implementation of the KL divergence-based regularizer, it was averaged over multiple DOFs. In our implementation, however, we introduced a regularization term with high DOFs intact, which was effective in diverse image generation.

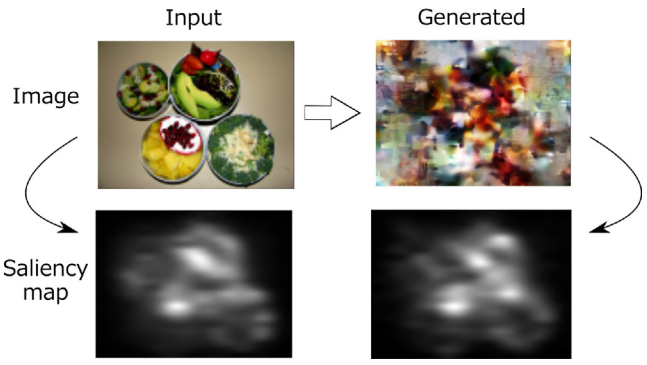

Fig. 1. Aim of the study. Our main objective is to develop a deep learning-based image transformation method to generate a deconstructed image from a given natural image, both with the consistent visual saliency map. In the upper panel, an example input image (left) and the corresponding generated image (right) are shown. In the lower, visual saliency maps of the images in the upper, each estimated by f_θ, are shown.

Fig. 1 exemplifies our image transformation results. To demonstrate the utility of the generated images by our deconstruction method, we performed human behavioral experiments including measurements of overt eye movements, to see if our method could generate artificial images that were unnatural enough but simultaneously associated with comparable eye movements. Moreover, we performed non-invasive brain imaging experiments with functional magnetic resonance imaging (fMRI), in which a different set of subjects passively viewed the images and we compared the brain activities during the natural and unnatural (generated) image presentation. The experimental results suggested that our DNN-based image transformation could be a new tool to elucidate the distinctive networks in the human visual system between bottom-up and top-down attention.

The major contributions of our study are as follows:

- A new concept of image transformation that transforms some features and maintains others. In particular, the transformed features are context and local structure, while the maintained feature are saliency maps. To realize this idea, a new loss function is introduced that encourages the image generator (deconstructor) to output an image whose saliency map is similar to that of the original image.
- Another methodological contribution is a device to enrich the local variability in terms of color contrasts and shapes by enlarging the latent space to a three-dimensional tensor. This latent space expansion allows the generator to produce a variety of fake images with different local colors and shapes.
- The applicability of our new methodology was examined in several experiments with human subjects. The behavioral experiments suggested that the deconstructed images evoked eye movements comparable to or even better than the original natural images, but the context was not perceived. The fMRI experiments suggested a possible dissociation of the functional basis between bottom-up and top-down attention.

The remaining parts are organized as follows. In Section 2, we introduce related works that are important for presenting our method. In Section 3, we propose our method of deep learning-based image deconstruction with maintained saliency map. Human experimental methods are also described. In Section 4, we evaluate our deconstructed images in a quantitative manner. In

addition, we show human behavioral experimental results, as well as human brain imaging results. Section 5 is devoted to conclusion of our work.

2. Preliminaries and related works

2.1. Saliency map

A saliency map is a heat map that visualizes the salient regions in images or movies, such to attract mainly bottom-up attention. There have been a number of studies to present the way to construct saliency maps (Harel et al., 2006; Itti et al., 1998; Pan et al., 2017). Among those, Itti and colleagues presented a computational model to obtain a saliency map, in which low-level image/movie features such as color, contrast, orientation, and motion direction are integrated (Itti et al., 1998). The usage of Gaussian pyramid with different spatial scales allowed their method to extract locally salient regions whose feature values are different from those of their surroundings. These basic operations correspond to the functions of simple cells that work as first-order filters and complex cells that work as second-order filters, which are assumed to consist of mammalian early visual systems (Anzai, Ohzawa, & Freeman, 1999; Ohzawa & Freeman, 1986). Since the saliency map well reproduces eye gaze distribution (fixation map) of humans, it has been recognized as representation of salient regions that attract eye gaze (Itti, 2005). In this study, we used the Itti's computation model registered in the GBVS toolbox (Harel et al., 2006), which exhibited good consistency with the fixation map (KL = 1.03), according to MIT300 benchmark (Bylinskii et al., 0000).

2.2. SALICON and SalGAN

SALICON is a large-scale dataset containing pairs of natural images of various scenes and their corresponding saliency maps annotated by humans (Jiang, Huang, Duan, & Zhao, 2015). Each saliency map was estimated based on trajectories of cursors operated by multiple human annotators registered in Amazon Mechanical Turk. The dataset was designed for machine learning competitions; it consisted of 10,000 pairs of training data, 5000 pairs of verification data, and 5000 natural images for testing (with no saliency map).

SalGAN is a GAN-based estimation method of saliency maps (Pan et al., 2017), which was trained using the SALICON dataset, and showed comparable performance with those by other model-based methods in the MIT300 benchmark (Bylinskii et al., 0000).

2.3. Stochastic image generation

Here, we introduce the existing approaches to diverse image generation or transformation.

2.3.1. Pix2Pix

Pix2Pix is a well-known method for image-to-image translation; that is, an image could be generated from an input label image (Isola et al., 2017). One typical application of Pix2Pix is to generate a painted image from an input image that only includes the outline of the image. This paining problem is ill-posed, because what color would be used for a texture surrounded by outline is not unique. Pix2Pix used two techniques to solve such an ill-posed image generation problem. One is to use Dropout to generate diversity in the image generation, and the other is to employ U-net-based generator to make the generated image to well maintain the structures of the input label image, in which multi-scale skip connections are effective.

2.3.2. Variational autoencoder

Variational autoencoder (VAE) is a stochastic generation model that introduces latent space to the intermediate representation (Kingma & Welling, 2013). It has an encoder–decoder architecture, in which the encoder reduces dimensionality into an intermediate representation that also includes a latent space, while the decoder reconstructs the input to compensate for the stochastic factor in the intermediate representation. To avoid excessive randomness, VAE encourages the posterior of the latent variable not to diverge from the standard normal distribution (with mean 0 and variance 1). This idea is good for stability in learning, whereas the stochastic image generation by the decoder may sometimes work harmfully for generating clear images.

2.3.3. Posterior collapse

A phenomenon called posterior collapse occurs especially when the KL regularization term works more powerfully than the primary objective like the minimization of reconstruction errors, so that the whole optimization process falls into a local optimum (He, Spokoyny, Neubig, & Berg-Kirkpatrick, 2019; Huang, Tan, Lacoste, & Courville, 2018). In such a local optimum, the posterior distribution of the latent variable $p(\boldsymbol{z}|\boldsymbol{x})$ becomes the same as the prior distribution $p(\boldsymbol{z})$, so that the latent variable $\boldsymbol{z}$ works as just a random number to obey the standard normal distribution, regardless of the input $\boldsymbol{x}$. This causes input information to be neglected when the decoder attempts to reconstruct the input. Because of this, balancing the regularization term and the objective term has been a handcraft issue.

Moreover, in the conventional implementation, the KL regularization term was averaged over the latent variables which may have multiple degrees of freedom (DOFs) and over training samples in mini-batches (Kingma & Welling, 2013). Since this averaged regularizer encourages the latent variables to take correlated values with each other, the image generation/transformation may lack diversity, leading to difficulty in decoder learning.

When the mean $\boldsymbol{\mu}$ has a large variation dependent on the input, and the variance $\boldsymbol{\sigma}^2$ is close to 0, in the latent space of the intermediate representation, the reconstruction error in the decoder would be small; in this extreme case, VAE just reduces to autoencoder (AE). The KL regularization term can be seen to work against this model reduction, and hence work to facilitate learning under a good tuning of the balancing hyperparameter (Bowman et al., 2015). Several studies have attempted to avoid the issue of posterior collapse, by, for example, carefully designing loss functions (Xu & Durrett, 2018), quantizing the latent space (van den Oord, Vinyals, et al., 2017), and so on.

2.3.4. VAEGAN

VAEGAN is a combination of VAE and GAN, each of which makes up for their respective weak points (Larsen, Sønderby, Larochelle, & Winther, 2015). GAN (Goodfellow et al., 2014; Radford, Metz, & Chintala, 2015) generates relatively clean images through indirect learning based on adversarial losses, in which a pair of networks, a generator and a discriminator, compete with each other, so that the latter works as the regularizer for the former. However, GAN is prone to lose diversity due to mode collapse and may suffer from instability due to the gradient elimination. This defect can be eased by combining GAN with VAE. On the other hand, blurring of generated images, which is a major issue of VAE, can be eased by GAN.

2.3.5. BicycleGAN

BicycleGAN is an image transformation method in which pre-transformed and post-transformed images are conditioned on their individual label images. The architecture of BicycleGAN consists of VAE, GAN, and LR (Latent Regressor) (Zhu et al., 2017).

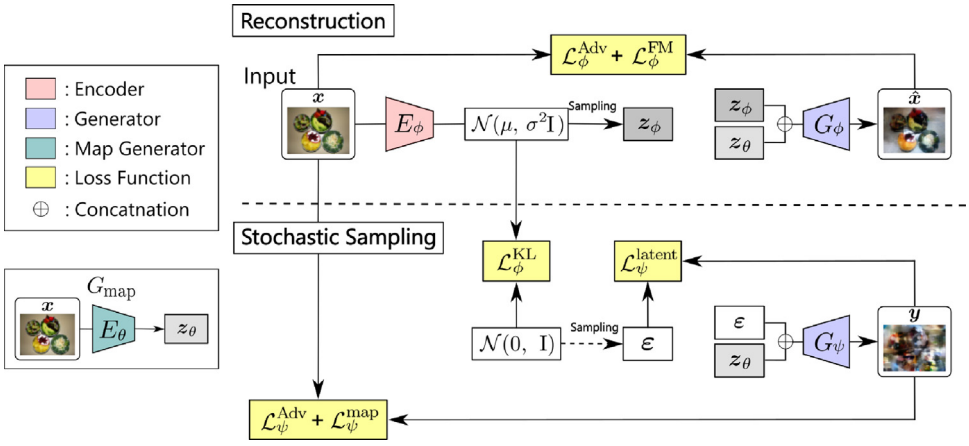

Fig. 2. Our architecture to generate deconstructed images with maintained saliency. It consists of three kinds of networks. The encoder network (red) is similar to that in VAE. The map generator network (green) is an encoder–decoder network to transform from an input image $\boldsymbol{x}$ to an output saliency map, whose intermediate representation is $\boldsymbol{z}_\theta$. There are two generators (purple), one is to generate a reconstructed input image, G_ϕ, and the other is to generate a transformed image, G_ψ. The two purple modules, G_ϕ and G_ψ, share their weights, i.e., they are the same. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

LR was introduced to maintain the bijection (one-to-one) relationship between two latent variables incorporated into a pair of VAEs. The loss function for the LR part was added to the loss functions of VAE and GAN. This combined loss function was expected to work well for generating diverse images from a pair of random numbers sampled in the dual latent space. BicycleGAN used VAE and GAN to generate realistic images, while LR was also incorporated to generate more diverse images. Consequently, this method exhibited excellent image transformation performance with good naturalness and diversity, whose balance-taking had been thought as of a typical dilemma.

Although all the above methods have shown excellent performance in the image-to-image transformation, they have mostly focused on transformation of styles or touches, which are global features over the whole image, whereas local image structures like shapes or colors have been maintained. Such an image transformation in the level of global features was realized by maintaining the local features; for example, BicycleGAN used skip-connections (Ronneberger et al., 2015) that transmit spatially local features extracted by the lower-layer networks to higher-layer networks. This technique was indeed effective in learning the global structures of images, though they cause a lack in local diversity in the generated images.

2.4. Techniques for improving image generation

We here describe a couple of techniques to improve the quality of image generation.

2.4.1. Feature matching

If there is a couple of networks, like in many image transformation methods, one simple technique to keep good correspondence between the two networks is to minimize the difference in the output between their corresponding intermediate layers (Salimans et al., 2016). Here, the intermediate layers are expected to represent image features. By measuring loss in the level of image features, instead of that in the level of input/output pixels, the image transformation or image discrimination can be robust against positional deviation, so that the generated images are less blurred.

2.4.2. Inception module

A conventional idea has been like: the deeper, the better, in the fields of DNN-based image processing. However, there is an opposite idea; that is, we can employ multiple and parallel small-sized network modules, called inception modules, each consisting of multiple convolution layers and a pooling layer, and an integration over the outputs from those modules becomes the whole-network output (Szegedy, Ioffe, Vanhoucke, & Alemi, 2017; Szegedy, Vanhoucke, Ioffe, Shlens, & Wojna, 2016). This makes it possible to enlarge the network effective size without increasing its depth, thus expecting enlargement in the image generation capacity while keeping efficiency in training the whole network.

3. Method

3.1. General learning scheme

Fig. 2 depicts the proposed architecture for image transformation. It consists of three major modules, saliency map generator, image reconstructor, and image deconstructor, which in total generate a reconstructed color image $\hat{\boldsymbol{x}} \in \mathbb{R}^{3\times H\times W}$ and a transformed color image $\boldsymbol{y} \in \mathbb{R}^{3\times H\times W}$, given an input color image $\boldsymbol{x} \in \mathbb{R}^{3\times H\times W}$, where we expect $\hat{\boldsymbol{x}}$ and $\boldsymbol{y}$ to have similar saliency maps. H and W denote the height and width of the input, reconstructed, and transformed images. This architecture enables us to obtain a diversely deconstructed but still clear image which has a similar saliency map to that of the input image.

The saliency map generator f_θ is a module that estimates its visual saliency map based on the input image $\boldsymbol{x}$ and is also used to estimate the visual saliency map of the stochastically generated image $\boldsymbol{y}$. Although there is a number of definitions of visual saliency map, we used the one presented by Itti and colleagues (Itti et al., 1998), because of its general recognition in the field of computational neuroscience. Fig. 3 shows its encoder–decoder architecture, and the whole network was trained by minimization of adversarial loss and feature matching loss, Eq. (1).

The middle layer representation $\boldsymbol{z}_\theta \in \mathbb{R}^{C\times H/2^2\times W/2^2}$ of the saliency map generator f_θ was expected to include essential features for generating saliency maps, and then used for augmenting the inputs to the generator modules, G_ϕ and G_ψ. The size of $\boldsymbol{z}_\theta$ was comparable to that of the latent variable in the image reconstructor, i.e., $\boldsymbol{z}_\phi \in \mathbb{R}^{2C\times H/2^2\times W/2^2}$, to allow it to have information sufficient for reconstructing the input image. C is the number of channels (in our particular implementation, $C = 8$).

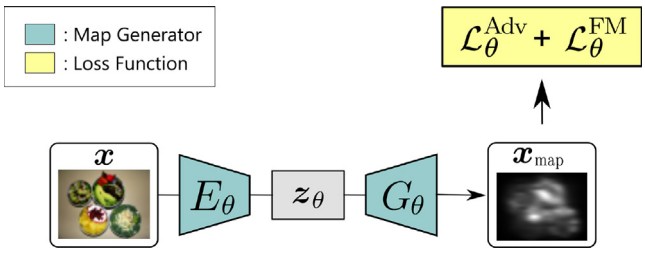

Fig. 3. Saliency map generator. This module outputs a visual saliency map, given an input image. It has been pre-trained based on the supervised dataset, consisting of pairs of a natural image taken from the SALICON dataset and the corresponding saliency map obtained by the method by Itti and colleagues (Itti et al., 1998). Note that the sole encoder part was used in the network in Fig. 2 for generating deconstructed images.

3.2. Training details

Here, we describe the loss functions used for training the three modules; saliency map generator $f_\theta = \{E_\theta, G_\theta\}$, image reconstructor $f_\phi = \{E_\phi, G_\phi\}$, and image deconstructor $f_\psi = \{G_\psi\}$. Note that the image generator is represented by a sole decoder G_ψ, which is actually the same as the decoder of the image reconstructor G_ϕ.

3.2.1. Saliency map generator

This module transforms an input image into its saliency map, and is trained to minimize the following loss function:

$$\mathcal{L}_\theta = \mathcal{L}_\theta^{\mathrm{FM}} + \lambda_\theta^{\mathrm{Adv}} \mathcal{L}_\theta^{\mathrm{Adv}}, \tag{1}$$

where $\lambda_\theta^{\mathrm{Adv}}$ is a pre-determined constant, a hyperparameter, to balance the two loss terms. They are defined by

$$\mathcal{L}_\theta^{\mathrm{FM}}(E_\theta, G_\theta, D_\theta) = \mathbb{E}\left[\sum_l \| D_\theta^l(\boldsymbol{x}, \boldsymbol{t}) - D_\theta^l\left(\boldsymbol{x}, \boldsymbol{x}_{\mathrm{map}}\right) \|_2\right], \tag{2}$$

$$\mathcal{L}_\theta^{\mathrm{Adv}}(E_\theta, G_\theta, D_\theta) = \mathbb{E}\left[-\log\left(D_\theta\left(\boldsymbol{x}, \boldsymbol{x}_{\mathrm{map}}\right)\right)\right], \tag{3}$$

where $\boldsymbol{x}$ and $\boldsymbol{t}$ are a natural image taken from SALICON and its corresponding saliency map calculated by the Itti's method (Itti et al., 1998), and $\boldsymbol{x}_{\mathrm{map}} \in \mathbb{R}^{1\times H\times W}$ is the output of the saliency map generator for $\boldsymbol{x}$, i.e., $\boldsymbol{x}_{\mathrm{map}} = f_\theta(\boldsymbol{x})$. $D_\theta\left(\boldsymbol{x}, \boldsymbol{x}_{\mathrm{map}}\right)$ denotes the output of the discriminator, which attempts to discriminate between a real pair ($\boldsymbol{x}$ and $\boldsymbol{t}$), whose ideal output is one, and a fake pair ($\boldsymbol{x}$ and $\boldsymbol{x}_{\mathrm{map}}$), whose ideal output is zero. D_θ^l denotes the feature vector of the lth layer of the discriminator D_θ. Note that the intermediate representation, $\boldsymbol{z}_\theta$, of f_θ was used as inputs for f_ϕ and f_ψ.

3.2.2. Image reconstructor

The loss function of the image reconstructor module f_ϕ is given by

$$\mathcal{L}_\phi = \mathcal{L}_\phi^{\mathrm{FM}} + \lambda^{\mathrm{KL}} \mathcal{L}_\phi^{\mathrm{KL}} + \lambda_\phi^{\mathrm{Adv}} \mathcal{L}_\phi^{\mathrm{Adv}}, \tag{4}$$

which consists of three terms, $\mathcal{L}_\phi^{\mathrm{FM}}$, $\mathcal{L}_\phi^{\mathrm{KL}}$ and $\mathcal{L}_\phi^{\mathrm{Adv}}$. λ^{KL} and $\lambda_\phi^{\mathrm{Adv}}$ are predetermined constants (hyperparameters).

The first term, $\mathcal{L}_\phi^{\mathrm{FM}}$, is the feature matching loss, given by

$$\mathcal{L}_\phi^{\mathrm{FM}}(E_\phi, G_\phi, D_\phi) = \mathbb{E}\left[\sum_l \| D_\phi^l(\boldsymbol{x}) - D_\phi^l\left(\hat{\boldsymbol{x}}\right) \|_2\right], \tag{5}$$

where D_ϕ^l denotes the feature vector of the lth layer of the discriminator D_ϕ, and $\hat{\boldsymbol{x}}$ is the output of the image reconstructor when $\boldsymbol{x}$ is input, i.e., $f_\phi(\boldsymbol{x}, \boldsymbol{z}_\theta)$. Here, the feature matching loss is used as a reconstruction error between the input image $\boldsymbol{x}$ and the reconstructed image $\hat{\boldsymbol{x}}$.

The second term, $\mathcal{L}_\phi^{\mathrm{KL}}$, is a regularization term for the latent variable, given below, which is different from the conventional KL regularization term.

$$\begin{aligned}\mathcal{L}_\phi^{\mathrm{KL}}(E_\phi) = {} & \mathcal{D}_{\mathrm{KL}}\left(\mathcal{N}(\mathbb{E}[\boldsymbol{\mu}_i],\ \mathbb{V}[\boldsymbol{\mu}_i] + \mathbb{E}[\boldsymbol{\sigma}_i^2])\ \|\ \mathcal{N}(0, \boldsymbol{I})\right) \\ & + \mathcal{D}_{\mathrm{KL}}\left(\mathcal{N}(\mathbb{E}[\boldsymbol{\sigma}_i^2],\ \mathbb{V}[\boldsymbol{\mu}_i])\ \|\ \mathcal{N}(\alpha, (1-\alpha)\boldsymbol{I})\right).\end{aligned} \tag{6}$$

Here, $\mathcal{N}(0,\ \boldsymbol{I})$ denotes a standard normal distribution. This standard normal prior has been used in many existing VAE-based image transformation (Zhu et al., 2017). Its benefit mainly comes from the calculation efficiency; the constraint loss in terms of the KL-divergence can be calculated in a closed form based on the mean and variance that are outputs of the VAE encoder. We can use a different distribution as a latent prior. In the case of general prior, however, it would be difficult to perform backpropagation-based training of the VAE encoder, because we need to evaluate the difference in the higher order statistics between the prior and the empirical distribution represented by the VAE encoder. To avoid such a difficulty, we simply introduced the standard normal prior to the VAE latent space. See Section 3.4 for more details.

The third term, $\mathcal{L}_\phi^{\mathrm{Adv}}$, is the adversarial loss, given by

$$\mathcal{L}_\phi^{\mathrm{Adv}}(E_\phi, G_\phi, D_\phi) = \mathbb{E}\left[-\log\left(D_\phi\left(\hat{\boldsymbol{x}}\right)\right)\right], \tag{7}$$

where $D_\phi\left(\hat{\boldsymbol{x}}\right)$ is the discriminator output when the reconstructed image $\hat{\boldsymbol{x}}$ is the input. Here, the ideal output of D_ϕ for an input $\hat{\boldsymbol{x}}$ is zero. This loss was effective in generating clearer images.

3.2.3. Image deconstructor

Since there is no ground truth for a deconstructed image, we trained the image deconstructor module, the sole generator, indirectly as to minimize the following loss function:

$$\mathcal{L}_\psi = \lambda^{\mathrm{latent}} \mathcal{L}_\psi^{\mathrm{latent}} + \lambda^{\mathrm{map}} \mathcal{L}_\psi^{\mathrm{map}} + \lambda_\psi^{\mathrm{Adv}} \mathcal{L}_\psi^{\mathrm{Adv}}, \tag{8}$$

where $\lambda^{\mathrm{latent}}$, λ^{map} and $\lambda_\psi^{\mathrm{Adv}}$ are the predetermined constants (hyperparameters). Note that $\lambda^{\mathrm{latent}}$ is not necessarily one, because we need to balance $\mathcal{L}_\phi$ (Eq. (4)) and $\mathcal{L}_\psi$ (Eq. (8)) when training the same generator, $G_\phi = G_\psi$.

The first term, $\mathcal{L}_\psi^{\mathrm{latent}}$, is the latent reconstruction error, given by

$$\mathcal{L}_\psi^{\mathrm{latent}}(E_\phi, G_\psi) = \mathbb{E}\left[\|\boldsymbol{\varepsilon} - E_\phi(\boldsymbol{y})\|_2\right], \tag{9}$$

where $E_\phi(\boldsymbol{y})$ is the output of the encoder of the image reconstructor for $\boldsymbol{y}$, which is in turn the output of the image deconstructor $G_\psi(\boldsymbol{\varepsilon}, \boldsymbol{z}_\theta)$. This term encourages the consistency of the latent variable $\boldsymbol{\varepsilon}$ when going through a network consisting of the generator G_ψ and E_ϕ, but alleviates mode collapse and helps a diversity of the image deconstructor. Although this loss function is defined by E_ϕ, only the parameters of the image deconstructor part G_ψ were updated based on Eq. (9), as in BicycleGAN.

The second term, $\mathcal{L}_\psi^{\mathrm{map}}$, is the loss between the saliency map for the input image $\boldsymbol{x}$ and that for the stochastically generated image $\boldsymbol{y}$, measured in terms of binary cross-entropy (BCE):

$$\begin{aligned}\mathcal{L}_\psi^{\mathrm{map}}(G_\psi, E_\theta, G_\theta) = \mathbb{E}[& -\{f_\theta(\boldsymbol{x}) \log(f_\theta(\boldsymbol{y})) \\ & + (1 - f_\theta(\boldsymbol{x})) \log(1 - f_\theta(\boldsymbol{y}))\}].\end{aligned} \tag{10}$$

The third term, $\mathcal{L}_\psi^{\mathrm{Adv}}$, is the adversarial loss, given by

$$\mathcal{L}_\psi^{\mathrm{Adv}}(G_\psi, D_\psi) = \mathbb{E}\left[-\log\left(D_\psi(\boldsymbol{y})\right)\right], \tag{11}$$

where $D_\psi(\boldsymbol{y})$ is the output of the discriminator D_ψ for the input $\boldsymbol{y}$; because $\boldsymbol{y}$ is a fake image, its ideal output would be zero.

3.2.4. Discriminators

In addition to the three major modules, we employed three discriminators for improving three generated images, $\boldsymbol{x}_{\text{map}}$, $\hat{\boldsymbol{x}}$ and $\boldsymbol{y}$ for a given input image $\boldsymbol{x}$. The individual discriminator was trained to discriminate if the input was a real image (or image pair) or a fake image (or image pair).

$$\mathcal{L}_{\theta}^{Adv} = \mathbb{E}\left[\log\left(D_{\theta}\left(\boldsymbol{x}, \boldsymbol{t}\right)\right)\right] + \mathbb{E}\left[\log\left(1 - D_{\theta}\left(\boldsymbol{x}, \boldsymbol{x}_{\text{map}}\right)\right)\right], \tag{12}$$

$$\mathcal{L}_{\phi}^{Adv} = \mathbb{E}\left[\log\left(D_{\phi}\left(\boldsymbol{x}\right)\right)\right] + \mathbb{E}\left[\log\left(1 - D_{\phi}\left(\hat{\boldsymbol{x}}\right)\right)\right], \tag{13}$$

$$\mathcal{L}_{\psi}^{Adv} = \mathbb{E}\left[\log\left(D_{\psi}\left(\boldsymbol{x}\right)\right)\right] + \mathbb{E}\left[\log\left(1 - D_{\psi}\left(\boldsymbol{y}\right)\right)\right]. \tag{14}$$

As in BicycleGAN, the discriminators D_{ϕ} and D_{ψ} are independent, and not conditioned by the label image (here, $\boldsymbol{t}$).

3.3. Latent matrix

In BicycleGAN, the latent variables were vectors of relatively small dimensionalities, so that they could include little positional information. This caused that latent variables tended to carry global information of the images like styles and touches, and then, the output images likely shared local information like contexts and local structures with the input images. As a result, it was difficult for generated images to have a wide variety, in which local shapes/colors and then context information had been largely changed. In this study, we made it easy for the latent variables to have positional information, changing the latent variables from vectors to feature maps of three-dimensional matrix (tensor), each of which is called a latent matrix. In addition, we did not use a fully-connected layer to convert a latent matrix into a one-dimensional vector; in the existing VAE learning, a fully-connected layer to convert a latent vector into another low-dimensional vector was employed (Zhu et al., 2017) and then positional information of the latent vector was diminished.

Since it is difficult to quantitatively evaluate unnaturalness of generated images, on the other hand, it is also difficult to set the loss function for the latent matrix. We then assumed that images with large complexity in its local shapes and colors are of high unnaturalness.

3.4. Processing for the minibatch of KL regularization terms

The increase in the component number of the loss function makes it difficult to balance by means of heuristic tuning of hyperparameters, often leading to the phenomenon of posterior collapse with a relatively strong KL regularization.

Let the latent variable obey a D-dimensional normal distribution:

$$\mathcal{N}(\boldsymbol{\mu}_i,\ \boldsymbol{\sigma}_i^2\boldsymbol{I}) \quad i = 1,\ 2,\ \ldots,\ N \tag{15}$$

which is assumed to be independent among samples in a minibatch. Here, N is the number of samples in this minibatch (in our implementation, $N = 10$), while $\boldsymbol{\mu}_i$ and $\boldsymbol{\sigma}_i^2$ are the mean and variance of the latent variable, respectively, for the ith sample. Note here that the mean and variance of each latent variable are assumed to be random variables conditioned on the input image, although they are in practice deterministic variables and outputs of the encoder neural network.

The conventional KL regularization term is given by

$$\mathbb{E}\left[\mathcal{D}_{\text{KL}}(\mathcal{N}(\boldsymbol{\mu}_i,\ \boldsymbol{\sigma}_i^2\boldsymbol{I})\ \|\ \mathcal{N}(0,\ \boldsymbol{I}))\right], \tag{16}$$

where $\mathbb{E}$ denotes the expectation over the empirical distribution in the minibatch.

It is obvious that the regularization based on Eq. (16) encourages all the latent variables in the minibatch to approach individually to the standard normal distribution; with poor hyperparameter tuning, it causes posterior collapse, leading to loss of information associated with the input image. To avoid this, we developed a new regularizer by considering the variance of the latent variables, according to the reparameterization trick (Kingma & Welling, 2013):

$$\begin{aligned}\mathbb{V}\left[\mathcal{N}(\boldsymbol{\mu}_i,\ \boldsymbol{\sigma}_i^2\boldsymbol{I})\right] &= \mathbb{E}\left[(\boldsymbol{\mu}_i\ +\ \boldsymbol{\varepsilon}_i \cdot \boldsymbol{\sigma}_i)^2\right] - \left(\mathbb{E}\left[(\boldsymbol{\mu}_i\ +\ \boldsymbol{\varepsilon}_i \cdot \boldsymbol{\sigma}_i)\right]\right)^2 \\ &= \mathbb{V}\left[\boldsymbol{\mu}_i\right] + \mathbb{E}\left[\boldsymbol{\sigma}_i^2\right],\end{aligned} \tag{17}$$

where $\boldsymbol{\varepsilon}_i$ is a D-dimensional random number sampled from the D-dimensional standard normal distribution, while $\mathbb{E}$ and $\mathbb{V}$ denote the expectation and variance over the empirical distribution of samples in the minibatch. Note that the left-hand side just denotes the empirical variance of the latent variables.

Since we have assumed that the latent variables obey normal distributions with different means and different variables, their marginal distribution over the minibatch becomes an N-component normal (Gaussian) mixture distribution. Under the simplification that this Gaussian mixture distribution can be approximated by a single-mode normal distribution, the KL regularization term becomes:

$$\mathcal{D}_{\text{KL}}\left(\mathcal{N}(\mathbb{E}[\boldsymbol{\mu}_i],\ \mathbb{V}[\boldsymbol{\mu}_i] + \mathbb{E}[\boldsymbol{\sigma}_i^2])\ \|\ \mathcal{N}(0,\ \boldsymbol{I})\right). \tag{18}$$

On the other hand, a too strong regularization based on Eq. (18) results in $\mathbb{V}[\boldsymbol{\mu}_i] \to 0,\ \mathbb{E}[\boldsymbol{\sigma}_i^2] \to 1$ or $\mathbb{V}[\boldsymbol{\mu}_i] \to 1,\ \mathbb{E}[\boldsymbol{\sigma}_i^2] \to 0$. To avoid such posterior collapses, we introduced a new regularization term, given by Eq. (19) with a pre-determined hyperparameter $0 < \alpha < 1$, which encourages the individual latent distributions not to overlap with each other.

Although the second term in Eq. (19) is just a heuristic, thanks to this term, there is no need to take a careful balance between the reconstruction error and the KL regularization term. As a consequence, the new regularization term has significantly reduced the time and effort for tuning the hyperparameter.

$$\begin{aligned}&\mathcal{D}_{\text{KL}}\left(\mathcal{N}(\mathbb{E}[\boldsymbol{\mu}_i],\ \mathbb{V}[\boldsymbol{\mu}_i] + \mathbb{E}[\boldsymbol{\sigma}_i^2])\ \|\ \mathcal{N}(0,\ \boldsymbol{I})\right) \\ &\quad + \mathcal{D}_{\text{KL}}\left(\mathcal{N}(\mathbb{E}[\boldsymbol{\sigma}_i^2],\ \mathbb{V}[\boldsymbol{\mu}_i])\ \|\ \mathcal{N}(\alpha,\ (1-\alpha)\boldsymbol{I})\right).\end{aligned} \tag{19}$$

In comparison to the well-established ridge (L2) regularization, we found the new regularization, Eq. (19), exhibited increased diversity and improved sharpness of the generated images.

3.5. Implementation details

The inception module was used for each layer of the networks. We used PatchGAN (Isola et al., 2017) for the discriminators D_{ϕ} and D_{ψ}, and also used a technique called minibatch standard deviation. The minibatch standard deviation (Karras, Aila, Laine, & Lehtinen, 2017) is a technique to introduce perturbations to the last layer of the discriminators, based on the standard deviation over the minibatch. Without this perturbation, discriminators can easily identify fake inputs, because the diversity of fake images is often collapsed leading to peaky distribution that is quite different from the real image distribution.

The saliency map generator f_{θ} was pre-trained using the supervised dataset of pairs of a natural image taken from the SALICON dataset and the corresponding saliency map obtained by the Itti's method (Itti et al., 1998), and used with the fixed weights when training the entire architecture for image deconstruction. All images, including those used for pretraining of the saliency map generator f_{θ}, were downscaled to the half in height and width, i.e., to the size of 96×128 (pixels). When training the saliency map generator, we used saliency maps of 96×128 (pixels) for supervised outputs; they were obtained by applying a downscaling after generating higher resolution saliency maps of 192×256 (pixels) for the original high resolution natural images according to the Itti's method (Itti et al., 1998). Each building

block was composed of 3 inception modules and 2 pooling layers; for example, it had five modules: the first inception module, the first up-sampling (down-sampling) block, the second inception module, the second up-sampling (down-sampling) block, and the third inception module. The last layer of each encoder reduced the dimensionality from 256 to 16, so the size of the latent variables $\boldsymbol{z}_\phi$ was set to $[N, 16, 24, 32]$. Here, N is the size of the minibatch ($N = 10$), and the size of the intermediate representation $\boldsymbol{z}_\theta$ of the saliency map generator f_θ was set to 8. Further details of the network architecture are presented in the Appendix A.

We set the hyperparameter values to $\lambda_\theta^{\mathrm{Adv}} = 0.1$, $\lambda_\phi^{\mathrm{KL}} = 10$, $\lambda_\phi^{\mathrm{Adv}} = 0.1$, $\lambda_\psi^{\mathrm{latent}} = 0.1$, $\lambda_\psi^{\mathrm{map}} = 0.001$ and $\lambda_\psi^{\mathrm{Adv}} = 0.01$. Since GAN sometimes causes loss in diversity due to the mode collapse, we set the hyperparameters so that the discriminator becomes slightly stronger than the generator. The minibatch size was set to 10 and the learning rate to 0.0003, and Adam (Kingma & Ba, 2014) was used to optimize the network parameters.

3.6. Human experiments

To demonstrate the utility of our image transformation-based methodology for investigating the neural mechanisms involved in top-down and bottom-up visual attentions in awake animals/humans, we conducted human behavioral and non-invasive brain imaging experiments.

All human experiments described in this section were done in accordance with the Declaration of Helsinki (World Medical Association, 1964) and approved by the two ethics committees of Graduate School of Informatics, Kyoto University (Kyoto, Japan) and Advanced Telecommunications Research Institute International (ATR) (Kyoto, Japan). The fMRI experiment was also approved by the safety committee of ATR. All subjects were volunteers, had normal or corrected-to-normal vision and provided written informed consent to participate in the experiment.

First, we conducted a human behavioral experiment of a visual discrimination task to examine the degree of naturalness of the original natural images and the artificially generated images. We obtained 10,000 pairs of natural images from the SALICON database and their corresponding transformed images with saliency map maintained, and then selected the top 400 image pairs with the smallest mean squared error (MSE) between the original natural image $\boldsymbol{x}$ and its reconstructed image $\hat{\boldsymbol{x}}$. These 400 pairs, 800 images in total, were used as experimental stimuli. Eight subjects observed each image on a computer display for three seconds and responded by pressing a computer key if they understood the context of the displayed image. Each subject performed two sessions consisting of 200 natural and 200 transformed/generated images. Any pair of the generated and the corresponding natural image was not presented in the same session, and the order of the images was randomized but common across the eight subjects.

Next, in a separate behavioral experiment, eye movements were measured while another set of human subjects viewed natural and generated images. Thirteen subjects (eight females, ages 20∼29), who had no ocular (color) dysfunctions and normal or corrected-to-normal sight (more than 0.5 vision) were participated. Each still image was presented for 4 s on a 23.6-inch monitor (Iiyama ProLite B2409HDS, Mouse Computer Co. Ltd., Tokyo, Japan), and the subjects watched it overtly at a distance of 0.6 m from the display in a dark room. A chin rest was used and the visual angle of each image was 32.5°. Eye movements were measured with 120 Hz by an LED-based eye tracker (Tobii Pro X3-120, Tobii Technology, Tokyo, Japan). Each subject participated in eight sessions, each of which consisted of 12 natural images, 12 generated images, 12 shuffled images, and 12 shuffled images whose spatial frequency amplitudes were maintained from those of the original natural images (hereafter referred to as 'amp' images). Each shuffled image was created by pixel-wise shuffling from the original natural image, but the color of each pixel was not changed, and each amp image was obtained by applying additional constraints so that the image was similarly shuffled but the amplitude of each spatial frequency band was maintained at that of the original natural image (Fourier transform was applied for normalization: the maximum and minimum amplitudes were 0 and 1, respectively). The order of presentation of the four categories of images was random but common across the subjects. When evaluating how well the saliency map of a specific image represents its associated eye movements, we first obtained the fixation density map (FDM) by applying kernel density estimation using non-isotropic Gaussian kernels to the histogram of eye gaze points of 13 subjects as they viewed the image, and then took the KL divergence between the FDM and the normalized saliency map (corresponding to a probabilistic density map).

To further demonstrate the utility of our image deconstruction method, a functional magnetic resonance imaging (fMRI) experiment was conducted to measure brain activity while subjects viewed the four categories of still images. As the image stimuli, we selected 96 image pairs from the 400 image pairs used in the first behavioral experiment, which had the smallest mean squared error (MSE) between the original natural image $\boldsymbol{x}$ and its reconstructed image $\hat{\boldsymbol{x}}$. Five healthy subjects (four females, ages $21 \sim 48$) participated in the experiment. At the beginning of each trial, a fixation cross was displayed for 2–4 s, followed by one of four different visual stimuli (natural, generated, shuffled, or amp) flashing at 2.5 Hz for 4 s Each subject performed 8 sessions, each session consisting of 12 natural, 12 generated, 12 shuffled, and 12 amp image trials and 4 test trials. The order of trials was random but common across the five subjects. The test trials were introduced to ensure that the subjects were engaged in the experiment; a natural image with an additional red cross was presented and subjects were instructed to press a button. Results showed that only one subject failed to respond to the test image twice (2/32, error rate: 6.3%), while the other four subjects responded to the test image 100% correctly.

Neuroimaging data was acquired with a 3T Siemens Prisma scanner (Siemens Healthcare GmbH, Erlangen, Germany) with the standard 64 channel phased array head coil. Whole-brain functional images were collected with a multiband echo EPI sequence (TR = 1000 ms, TE = 30 ms, flip angle = 50, field of view = 100 mm), and 66 slices (voxel size $2 \times 2 \times 2.5$ mm) were acquired per volume. High resolution T1-weighted structural images (TR = 2250 ms, TE = 3.06 ms, voxel size $1 \times 1 \times 1$ mm) using standard MPRAGE sequence were also obtained.

All imaging analyses were conducted using SPM12 (Welcome Department of Cognitive Neurology, UCL, London, UK) in MATLAB (MathWorks Inc., Natick, USA). For preprocessing, all functional images were realigned and resliced to the reference functional volume, coregistered to the individual high-resolution anatomical image, spatially normalized to the standard East Asian Brain template with a resample voxel size of $2 \times 2 \times 2$ mm, and spatially smoothed with a Gaussian kernel filter (FWHM, 8 mm). After preprocessing, we conducted statistical imaging analysis using a generalized linear model (GLM).

We examined BOLD signals when the subjects were observing the natural and three kinds of unnatural images using a boxcar regressor for each condition, aligned to the onsets of the image presentation trials with 4 s. duration. Although the unnatural condition included the generated, shuffled, and shuffled with the maintained spatial frequency amplitude (amp) images, we integrated the last two classes, because there was no characteristic difference in evoked fMRI activities between the two classes. Each

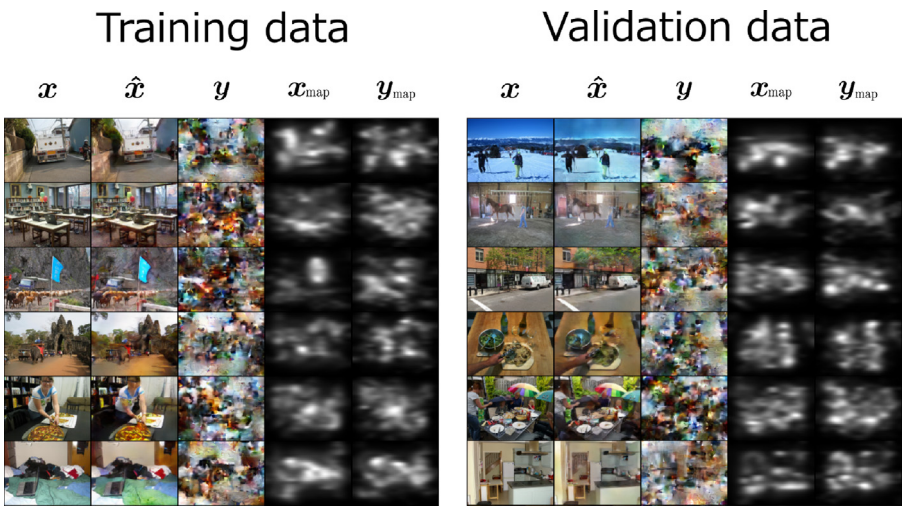

Fig. 4. Images generated by the proposed method. The left and right panels are for different input images; images used for training the image deconstructor (left) and those not used for the training (right). On each row, the images are as follows: the input natural image $\boldsymbol{x}$, the reconstructed image $\hat{\boldsymbol{x}}$, the deconstructed image $\boldsymbol{y}$ which is the result of the transformation to an unnatural image while maintaining the visual saliency, the saliency map of the input image $\boldsymbol{x}_{\text{map}}$, and the saliency map of the deconstructed image $\boldsymbol{y}_{\text{map}}$, from the left column to the right column.

Fig. 5. Diversity in the deconstructed image. Even when the same input was provided to the trained network (the left panel, a pair of a natural image and its saliency map), it could generate a variety of deconstructed images (the right panel, they well maintained the input saliency map).

boxcar function was then convolved with the canonical hemodynamic response function and entered as an orthogonalized regressor into a standard generalized linear convolution model. The six motion parameters produced during realignment were also used as regressors in the imaging analysis to account for residual effects of scan-to-scan head motions. Due to the small sample size, we applied a multi-subject conjunction analysis (FWE, $p < 0.05$) to localize the brain voxels that were distinctively activated, commonly over the subjects, during the observations between the natural, generated, and the other (shuffled + amp) images.

4. Results

4.1. Quality of the generated images

Fig. 4 presents images generated by the proposed method. The reconstructed images $\hat{\boldsymbol{x}}$ in the second column well reconstructed the input image $\boldsymbol{x}$ overall, while some details were distorted. In the deconstructed images $\boldsymbol{y}$ in the third column, on the other hand, local shapes and local colors in the input images were well collapsed, while keeping their saliency maps $\boldsymbol{y}_{\text{map}}$ in the fifth column similar to the original ones $\boldsymbol{x}_{\text{map}}$ shown in the fourth column. These observations were consistent regardless of the input image being in the training data or in the validation data.

Fig. 5 demonstrates the diversity in the deconstructed images produced by our method. Since the latent distribution is close to the standard normal distribution, a variety of deconstructed images could have been obtained by the latent variable with its sufficient stochasticity. The global arrangement of the objects agreed with that of the input image, whereas the local structures and colors of the deconstructed images were fairly distant from those of the input image, which was preferable for our objective.

The local colors were blurred and cluttered. This would have occurred because of the relatively large DOFs of the latent space we used. Since the mapping from the original colored object space to the color-based saliency space is many-to-one, there was an innumerable number of possible local color assignments to produce the same color-based saliency map. This character is also preferable for us because the object color is one of the most informative features for understanding the context of the whole image.

Because the performance of our image deconstructor depends on that of the saliency map generator (SMG), generalization capability of the SMG not only for natural images but also for generated (deconstructed) images was examined. Table 1 shows the results. Although the cosine similarity (CS) between the saliency map predicted by our SMG and that by the Itti's method with the

Table 1
Cosine similarity (CS) between the saliency map predicted by our saliency map generator (SMG) and those obtained by the Itti's method. We examined the Itti's saliency map with two different spatial resolutions; one is of high spatial resolution of 192 × 256, and the other is low spatial resolution of 96 × 128. We used 5000 natural images registered in the validation dataset in SALICON. Since this matrix of CS is symmetric and the diagonal elements should be the unity, we signified unnecessary entries by the mark –.

	SMG	Itti(192 × 256)	Itti(96 × 128)
SMG	–	0.960 ± 0.019	0.898 ± 0.050
Itti(192 × 256)	–	–	0.921 ± 0.042
Itti(96 × 128)	–	–	–

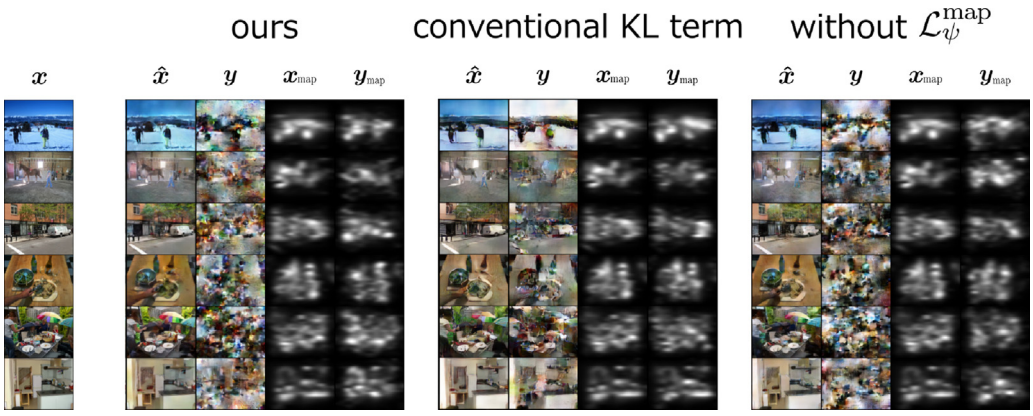

Fig. 6. Generated images after applying ablations to the proposed method. Input images (left-most panel), generated images by our proposed method ('ours', second-left panel), our method but with the conventional KL regularization term (Eq. (19)) ('conventional KL term', second-right panel), and our method but trained without the map loss function (Eq. (10)) ('without $\mathcal{L}_{\psi}^{\mathrm{map}}$', right-most panel).

Table 2
Fréche Inception Distance (FID) and Cosine Similarity (CS) between saliency maps. Each FID score represents the one over 5000 test images. Note that those test images were not used for training each of the three image transformation/generation methods. In the case of the CS score, the mean and standard deviation (SD) over 5000 test images are shown. Since the FID score is given for each method, there is no SD shown. The results for a lower image resolution of 48 × 64 are also shown. In the lower spatial resolution case, our method's CS was significantly improved over those of its ablated ones, while in the higher spatial resolution case of 96 × 128, its improvement over that without the map loss function was incremental.

Model	FID		CS	
	48 × 64	96 × 128	48 × 64	96 × 128
Ours	269.16	322.95	0.931 ± 0.020	0.908 ± 0.029
Conventional KL term	125.46	176.00	0.968 ± 0.012	0.950 ± 0.019
Without $\mathcal{L}_{\psi}^{\mathrm{map}}$	264.19	319.11	0.913 ± 0.026	0.903 ± 0.030

low resolution of 96 × 128 (pixels) was slightly lower, the general concordance between the saliency map predicted by our SMG and that by the Itti's method was satisfactorily high, suggesting the reliable generalization of the SMG. The relatively low CS between the SMG-based saliency map and that of the Itti's with the low resolution occurred, possibly because of the underlying resolution discrepancy between the two saliency maps. Actually, the size of the SMG output was 96 × 128, while the Itti's low-resolution saliency map was obtained by downscaling after getting a high-resolution saliency map by inputting a high-resolution natural image. Note that we used images and saliency maps of the consistent resolution of 96 × 128 (pixels) in this study; this was due to the restriction of the computer resources.

To evaluate the efficiency of our method, we measured the computational time required for the image deconstructor to learn a single input image in terms of FLOPs (number of floating-point operations). In our case of the input image size of 96 × 126, the saliency map generator, encoder, decoder and the discriminator, required 8.4 GFLOPs, 3.85 GFLOPs, 4.60 GFLOPs, and 0.97 GFLOPs, respectively. When we used NVIDIA DGX A100 (40 GB), it took approximately 8 h for learning with 30 epochs and approximately 5 mins to transform 10,000 natural images by the trained network.

There are two important components in our proposed method; one is the newly presented KL regularization term $\mathcal{L}_{\phi}^{\mathrm{KL}}$ (Eq. (19), conventional methods used Eq. (16)), and the other is the map loss function $\mathcal{L}_{\psi}^{\mathrm{map}}$ (Eq. (10)) used to keep the saliency to be maintained in the generated image. Here, we performed an ablation study to examine how well the two components above worked.

Fig. 6 displays the generated images by our proposed method, our method but with the conventional KL regularization term, and our method but trained without the map loss function. When we used the conventional KL regularization (Eq. (16)) instead of our newly developed KL regularization (Eq. (19)), the generated images $\boldsymbol{y}$ became similar to the reconstructed ones $\hat{\boldsymbol{x}}$, probably because the image deconstructor emphasized on the encoded features of the input images, $\boldsymbol{z}_{\theta}$, much more on the stochastic features $\boldsymbol{z}$. This is a typical example of posterior collapse; the regularizer was too strong to generate diverse images. If we remove the map loss function (Eq. (10)) from the loss function of the image deconstructor (Eq. (8)), the generated images were fairly diverse and unnatural, but their saliency maps were a little apart from those of the input images (Fig. 6 (right-most)). In contrast to those, our proposed method generated diverse and unnatural images, whereas their saliency maps were consistent with those of the input images (Fig. 6 (second-left)). Such favorable characters were owing to the newly developed KL regularization term, which was also effective in avoiding local optima in the training of the image deconstructor, and the map loss to keep the saliency map not to be much changed.

These visual inspections were further supported in a quantified manner. Table 2 shows the Fréche Inception Distance (FID), which has been used for evaluating the naturalness of the generated images in the field of image transformation/generation (Heusel, Ramsauer, Unterthiner, Nessler, & Hochreiter, 2017), and Cosine Similarity (CS) of a pair of saliency maps between the input image and the deconstructed image. A low (high) FID score suggests the naturalness (unnaturalness) of the transformed/generated images in terms of the distribution difference from the set of input images. A low (high) CS score suggests that the saliency map of the input image $\boldsymbol{x}_{map}$ and the one of the transformed/generated image $\boldsymbol{y}_{map}$ are quite different (similar). Our image deconstructor incorporated several techniques developed in the field of natural image transformation/recognition, such as feature matching, inception module, PatchGAN, and mini-batch standard deviation. Further ablation study supported their effectiveness in generating unnatural images with high variability (Appendix D).

To further examine the image deconstruction performance of our method, we compared our method with a couple of existing image transformation methods, Pix2Pix [1] and AdaIN [2]. Although AdaIN is an image style transformation method, and hence cannot deal in principle with ill-conditioned problem like our image deconstruction, we examined a modified version of AdaIN by attaching an additional loss function to facilitate the maintenance of the saliency map; this additional loss is given as $\mathcal{L}_{\psi}^{\mathrm{map}}$, being the same as the one used in our method.

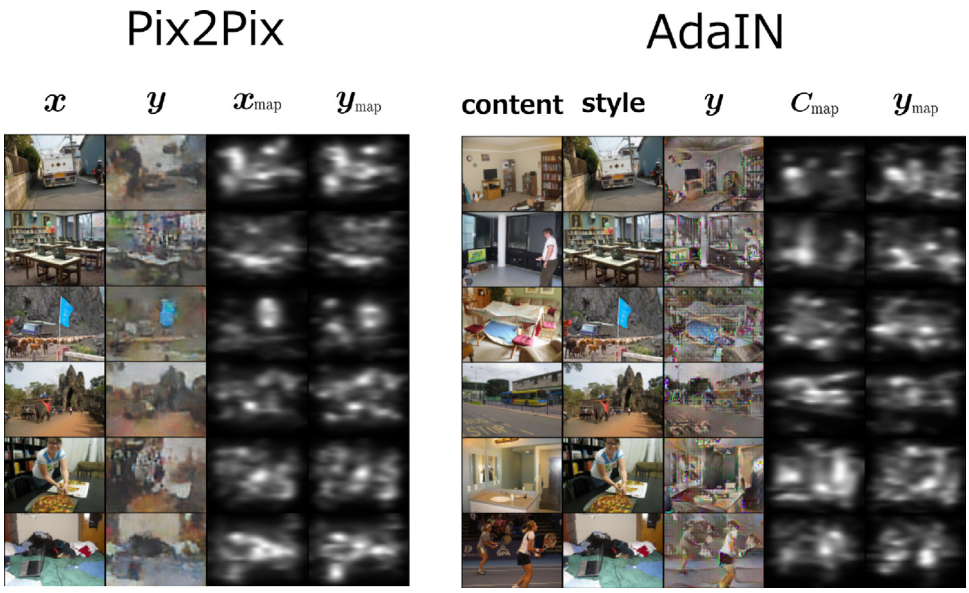

Fig. 7. Images generated by a couple of existing methods. These results show when testing the Pix2Pix (left panels) and AdaIN (right panels) methods after trained with the same image dataset as in our image deconstruction method. In the left panel, an original natural image, an image generated by Pix2Pix, the saliency map of the original image, and the saliency map of the generated image, are shown from the left to the right on each column. In the right panel, a natural, content image, a natural style image, an image generated by AdaIN, the saliency map of the input content image, and the saliency map of the generated image, are shown from the left to the right on each column.

Table 3
Cosine Similarity (CS) between the saliency map of an input natural image and that of a generated image. This table shows the saliency map consistency between the input image (or the content image in the case of AdaIN) and the output image.

Model	CS
Ours	0.908 ± 0.029
Pix2Pix	0.911 ± 0.028
AdaIN	0.890 ± 0.037

We trained Pix2Pix to transform an input saliency map to an output natural image, based on a training dataset of pairs of natural images (in SALICON) and the corresponding saliency maps; the latter was obtained according to the Itti's method. Fig. 7(left) shows the artificial images generated by Pix2Pix, which were found to be likely burred but still natural-like images that corresponded to the input saliency maps. The generator in Pix2Pix was of the U-net architecture with multi-resolution skip connections, while the image generation process from a saliency map to an artificial image would not have necessarily required such a shape-reserving transformation. Actually, the images generated by Pix2Pix maintained the global/local structures from those of the original natural images; this was different from what we expected.

In our implementation of AdaIN, we chose a couple of natural images randomly from the natural image set (i.e., SALICON), and set either one as a content image and the other as a style image. The objective of the AdaIN training is to transform the content image into the one of the style image-like style. The images generated by AdaIN maintained the global structures of the content images, but their styles looked similar to those of the style images (Fig. 7(right)).

Table 3 shows the Cosine Similarity (CS) between the saliency map of the input (in the case of Pix2Pix) or the content image (in the case of AdaIN), and that of the generated image (in both cases). We see that the saliency map of the Pix2Pix-based generated image was more similar to the input saliency map than that by our method; this was natural, because the objective of the Pix2Pix training was to perform the inverse process of the generation of saliency maps from natural images. However, the generated image by Pix2Pix was much similar to the original natural image, which was apart from our purpose. On the other hand, the saliency map of the AdaIN-based generated image less maintained than that by our method, though the AdaIN and our method used the same loss function to maintain the saliency map.

Appendix E describes further details of the implementations of Pix2Pix and AdaIN used here.

These results suggested that our proposed method was effective in generating diverse deconstructed/unnatural images while keeping their saliency maps similar to those of the corresponding input images.

4.2. Behavioral experiment results

In the visual discrimination task, each subject was asked to answer whether the presented image was natural or not in its context. Fig. 8 shows the results of the visual discrimination task. The fairly high accuracy for the 400 image pairs (97.9 ± 5.3%, SD is over the eight subjects) suggests that the original images taken from SALICON was deemed natural enough, whereas our generated images were deemed unnatural enough. Fig. 9 shows examples of generated images with a high accuracy rate (judged to be unnatural) in the behavioral experiment.

Next, we measured eye movements during thirteen human subjects were looking at natural, generated, shuffled, and shuffled but maintained spatial frequency amplitudes (amp) images. Fig. 10 shows the KL divergence, averaged over the 13 subjects, between the FDMs and the saliency maps. As there was no characteristic difference in eye movements between for the shuffled and for the amp images, they were merged together. The similarity between the FDM and the Itti's saliency map was highest when the generated images were presented, which was significantly better than when looking at natural images. We consider that this difference may be due to the fact that in unnatural situations where the generated images are presented, the eye gaze is mainly directed to bottom-up attentive regions in the images that match the Itti's saliency map well, whereas in natural

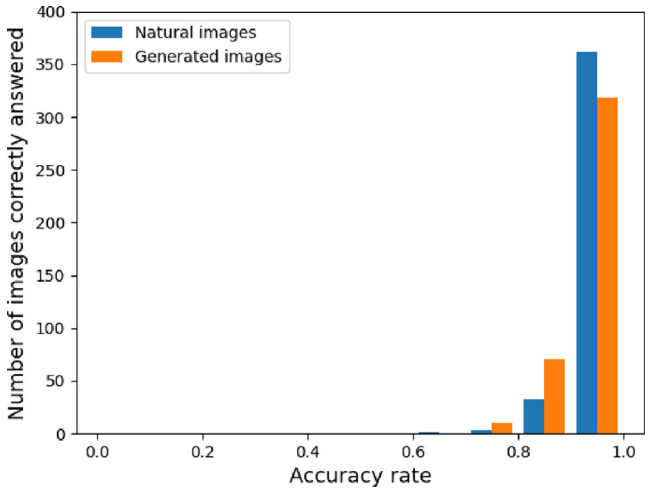

Fig. 8. Histogram of accuracy for 400 image pairs consisting of natural and generated images. If each subject answered natural for the original image taken from SALICON database or unnatural for the generated image, the answer was considered correct, otherwise incorrect. In the actual experiment, subjects were asked whether they could understand the context of each image shown on the computer display. The horizontal axis is the rate of correct answers of the eight subjects (accuracy rate) and the vertical axis is the number of images for which the accuracy rate (or correct response rate) was the value on the horizontal axis (blue: original natural images, orange: generated images). Note that the values on the horizontal axis are discrete: 1/8, 2/8, 3/8, 4/8, 5/8, 6/8, 7/8 and 8/8. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

situations the eye gaze is also attracted to top-down attentional regions that are not sufficiently represented by the Itti's saliency map. We also found that the FDMs when viewing shuffled images were closer to the saliency maps than when viewing natural images. This is because the saliency map is uniform in shuffled images, and the FDM is distributed around the center of the image because, by default, subjects tended to gaze near the image center when presented images without obvious saliency. In fact, the KL distance to the FDMs with shuffled images became the smallest with a single-mode Gaussian density centered on the image's center; in this case, the Gaussian density was estimated based on all the FDMs of the four image categories.

4.3. fMRI analysis results

We performed differential analysis of brain activities when the subjects observed the natural images and the images generated by our image deconstructor (Fig. 11). When observing the natural images, bilateral extrastriate visual area (BA7/19/37) showed significantly higher activities, but not the primary visual cortex (Fig. 11(a)). In contrast, bilateral primary visual cortices (BA17/18) and bilateral inferior parietal lobules (BA39/40) showed greater activities when viewing generated images (Fig. 11(b)).

We also compared the brain activities while viewing the natural, and the shuffled and amp images (images with preserved spatial frequency amplitudes). The latter two kinds of images are as unnatural as the generated images, but do not have the same saliency maps as the original natural images. The results of the differential analysis showed no prominent differences between natural vs. generated and natural vs. shuffled (+ amp) (Fig. 12), suggesting that the contextual information was collapsed in the generated images as well as in the shuffled and amp images.

In addition, brain activity was compared while the subjects were observing the generated and the shuffled/amp images (Fig. 13). Note that they shared the same color histogram. When the subjects observed the generated images, bilateral extrastriate cortices (BA18/19) showed significantly higher activation than when they observed the shuffled/amp images.

5. Discussions

5.1. Limitation in the methodology

The diversity of the generated deconstructed images was due to the distribution of the latent variable. To this end, we applied a regularization term to facilitate the distribution of all the latent variables to approach the standard normal distribution. When we examined the latent distribution, we found it had larger kurtosis than that of normal distribution, while its expectation and variance were close to 0 and 1, respectively. This has arisen because we encouraged the latent distribution over a minibatch, which should be a mixture Gaussian distribution, to approach to a unimodal Gaussian distribution. This may be improved by considering the KL divergence under an assumption that the target latent distribution is approximated by a multi-modal distribution like mixture Gaussian with several components.

We also examined a technique to add the adversarial loss which makes all the latent distributions individually close to normal distributions (Makhzani, Shlens, Jaitly, Goodfellow, & Frey, 2015; Mescheder, Nowozin, & Geiger, 2017; Rosca, Lakshminarayanan, Warde-Farley, & Mohamed, 2017). Although we found that the latent distributions well approached the standard normal distributions, there arose posterior collapse by ignoring the input information in the posterior latent distribution. Because of this, we have given up to proceed to this direction.

Since the latent distribution is important for the diversity of the generated images, the development of further sophisticated regularizer remains as a future study.

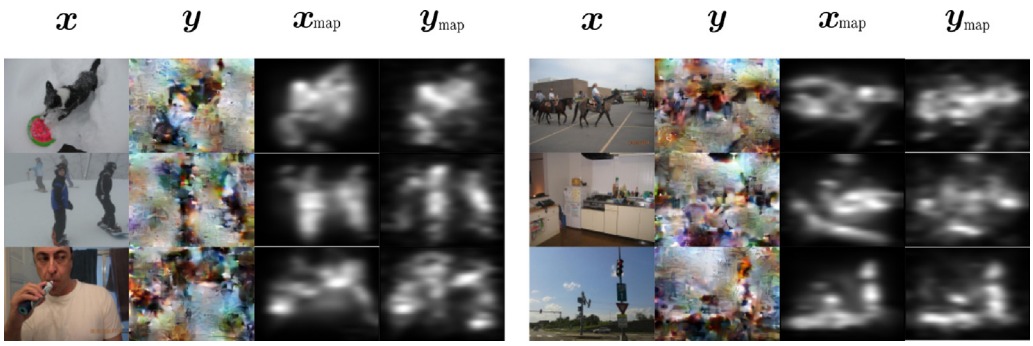

Fig. 9. Generated unnatural images. The generated image $\boldsymbol{y}$ were considered as sufficiently unnatural in the behavioral experiment, whereas they shared very consistent visual saliency maps with the original images $\boldsymbol{x}$.

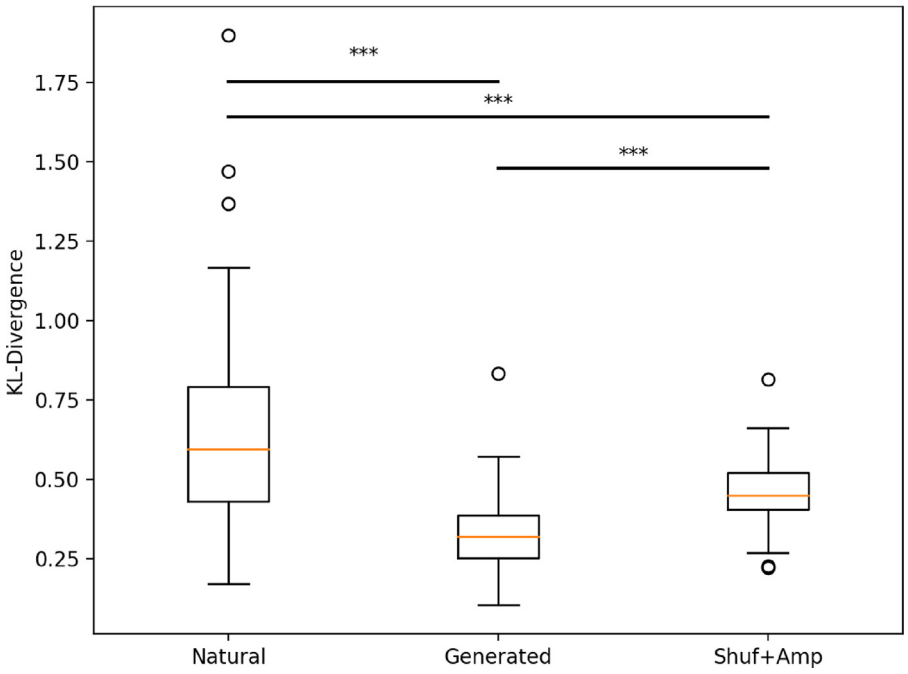

Fig. 10. The difference between the fixation density map (FDM) and the saliency map. Box-plots of KL divergence between the FDM based on eye movements of 13 subjects and the saliency map calculated by the Itti's method. Each box extends from the lower to upper quartiles, with a horizontal line at the median. The whiskers show 1.5×IQR, and cross markers indicate the outliers. The presented images were categorized natural, generated, and others (shuffled and shuffled but maintaining spatial frequency amplitude (amp)). Student t test showed that the FDM for the generated images was significantly more similar to the Itti's saliency map than the other images' FDM (natural: $p = 1.26e^{-15}$, others: $p = 5.37e^{-7}$) and the FDM for the shuffled images was significantly more similar to the saliency map than the FDM for the natural images ($p = 1.12e^{-12}$).

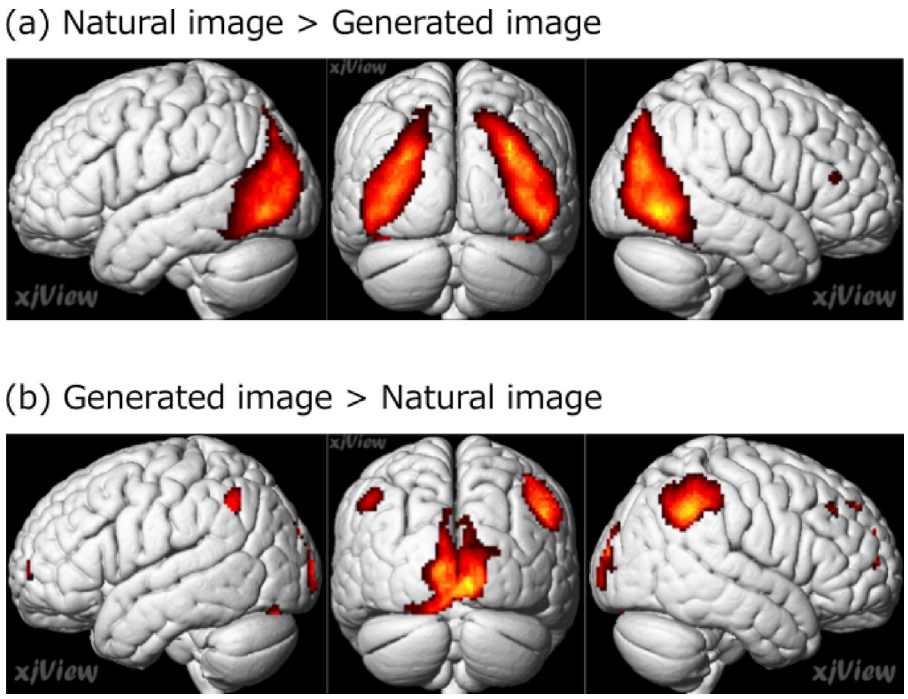

Fig. 11. Differential brain activity analysis between viewing the natural images and viewing the generated images. While the subjects observed the natural images, the bilateral extrastriatal cortices (Broadmann areas (BAs) 7/19/37, left: MNI coordinate = −44, −74, −6, right: 55, −66, −6) showed significantly higher activations (panel (a)). During the artificial image viewing, in contrast, the bilateral primary visual cortices (BA17/18, left: −6, −84, −12, right: 10, −78, −6), the bilateral inferior parietal lobules (BA39/40, left:,−48, −58, 46, right: 50, −50, 42) showed greater activities (panel (b)). The results are based on a multi-subject conjunction analysis (FWE, $p < 0.05$) and visualized using xjView toolbox (https://www.alivelearn.net/xjview).

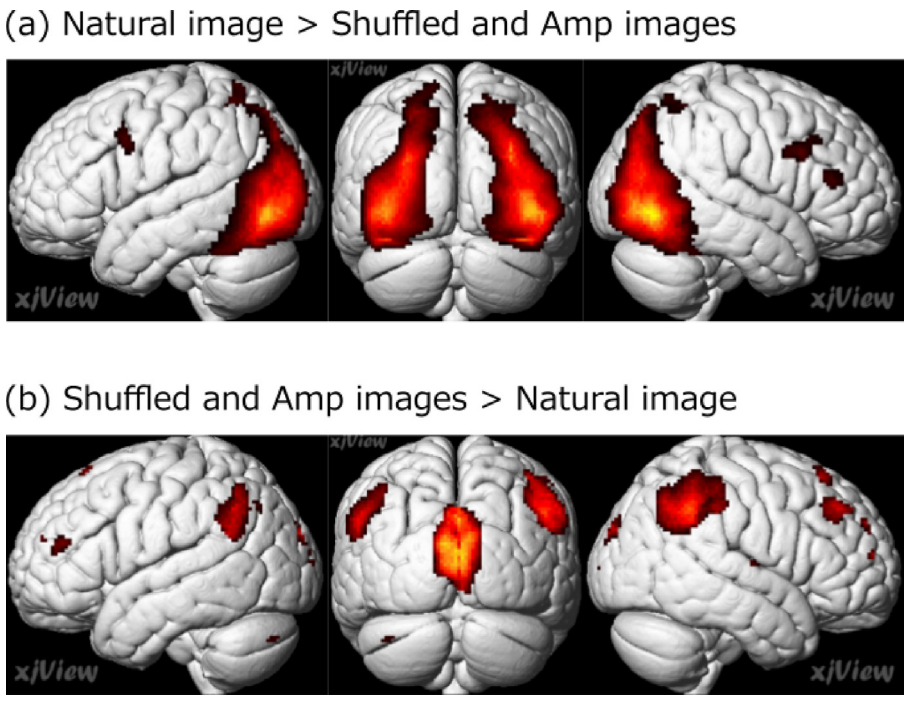

Fig. 12. Differential brain activity analysis between viewing the natural images and viewing the shuffled images. While the subjects observed the natural images, the bilateral extrastriatal cortices (Broadmann areas (BAs) 7/19/37, left: MNI coordinate $= -44, 76, -6$, right: $44, -76, -2$), and the bilateral dorsolateral frontal cortices (BA9, left: $-48, 6, 38$, right: 40,14,28) showed significantly higher activations (panel (a)). During the shuffled image viewing, in contrast, the bilateral primary visual cortices (BA17/18, left: $-6, -90, 20$, right: $6, -86, 16$), the bilateral inferior parietal lobules (BA39/40, left:,$-56, -58, 36$, right: $54, -50, 44$) showed greater activities. The results are based on a multi-subject conjunction analysis (FWE, $p < 0.05$) and visualized using the xjView toolbox.

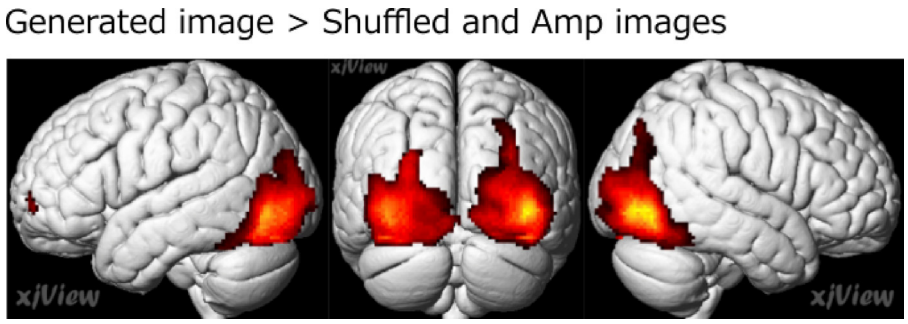

Fig. 13. Differential brain activity analysis between viewing the generated images and viewing the shuffled images. While the subjects observed the natural images, the bilateral extrastriatal cortices (Broadmann areas (BAs) 18/19, left: MNI coordinate $= -42, -76, -6$, right: $44, -78, -4$) showed significantly higher activations.

5.2. Neuroscientific implications

In this study, we conducted two kinds of behavioral experiments. The results from the visual discrimination task of image context suggested that our image deconstructor produced collapsed images with loss of context information. The results from eye movement measurement experiments also indicated that our image deconstructor retains the saliency map of the original image well, as the fixation density map (FDM) representing the eye movements corresponded best to the Itti's saliency map. In contrast, the FDM for the natural images did not resemble the Itti's saliency map. These results support the hypothesis that the Itti's saliency map mainly represents bottom-up attentional regions, but that eye movements are also modified by top-down attention, especially for the natural images. Based on this hypothesis, the images artificially generated by our image deconstructor could be used to dissociate the neural activity involved in bottom-up attention, even in overt image observation environments.

Although the main objective of this study is to present a new tool for computational neuroscientists to investigate the visual attention system, we conducted our own human non-invasive imaging experiments to demonstrate the usefulness of our tool. The results of the fMRI experiments showed that the brain activities while viewing the natural and the shuffled images were similar to those seen when comparing the natural and generated images, suggesting that the generated images lost context information that could be recognized by humans. The results of the differential analysis of brain activities for the two types of unnatural images, the generated and the shuffled image, showed significantly stronger activity in the middle-level visual cortices including V2 and V3, when the generated images were observed. This result is interesting because both the generated and shuffled images are unnatural, but with a saliency map that is not uniform for the former but almost uniform for the latter. Although the present imaging study was preliminary and these results require further evaluation studies, we believe that our proposed method

Table A.4
Detailed architectures of the sub-networks used in our image deconstructor. Block(C_{in}, C_{out}, k, activation) means a single block (Fig. A.14(a)), where C_{in}, C_{out}, and k mean the sizes of input channel, output channel, and kernel, respectively, and activation signifies the activation function. 'Down' and 'Up' mean a downsample and an upsample module (Fig. A.14(b)(c)), respectively, and 'Inc(C)' means an inception module of its input/output channel size being C. 'Conv(C_{in}, C_{out}, k)' means a block from which the batch normalization and the activation function were removed from a normal Block module. 'DeConv(C_{in}, C_{out}, k)' is a deconvolution block with a sigmoidal activation function. The deconvolution block used in the 'Map Decoder' and 'Decoder' employed convolution operations that were transformed from those in the corresponding convolution Block.

Map generator	Encoder	Decoder	Discriminator
Block(3, 64, 1, L)	Block(3, 64, 1, L)	Block(24, 256, 1, L)	Block(3, 32, 1, L)
Inc(64), Down	Inc(64), Down	Inc(256), Up	Inc(32), Down
Inc(128), Down	Inc(128), Down	Inc(128), Up	Inc(64), Down
Inc(256)	Inc(256)	Inc(64)	Inc(128)
Block(256, 8, 1, L)	Conv(256, 16, 1)	DeConv(64, 3, 1), Sigmoid	DeConv(129, 1, 1), Sigmoid
Block(8, 256, 1, L)			
Inc(256), Up			
Inc(128), Up			
Inc(64)			
DeConv(64, 1, 1), Sigmoid			

may provide a novel experimental tool to dissociate the neural bases involved in bottom-up and top-down attention.

6. Conclusion

In this study, we presented a new deep learning-based image transformation method with a relatively large latent space, which output deconstructed images based on the input natural images with maintaining the saliency maps of the input images. As a new latent regularization, KL regularization was proposed for the distribution of all latent variables, avoiding the collapse of the latent distribution and stabilizing the learning process.

The results of the behavioral analyses well validated that our new image transformation is effective in destroying contextual information embedded in natural images while preserving local structures. In addition, the results of fMRI experimental study suggested that different brain networks were evoked between when natural images were presented and when deconstructed images with the consistent saliency maps were presented.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This study was supported by Grant-in-Aid for Scientific Research on Innovative Areas, MEXT KAKENHI (JP17H06310), Japan, Grant-in-Aid for Scientific Research, JSPS KAKENHI (JP19H04180, JP22H04998), Japan, New Energy and Industrial Technology Development Organization (NEDO), Japan, and The Brain Mapping by Integrated Neurotechnologies for Disease Studies (Brain/MINDS) from AMED, Japan.

Appendix A. Detailed network architectures of the proposed method

First, we present specifications of the modules used in our DNN-based image deconstructor (Fig. A.14). A Block was a convolution layer employing a leaky ReLU or no activation function (panel (a)). In a Downsample (an Upsample) module, the channel size C, image height H, and image width W were halved (doubled) between the input and output (panel (b) and (c), respectively). We did not change the size of channel, image height, or image width, in our inception modules (panel (d)).

Table A.4 shows the detailed architectures of the sub-networks used in our image deconstructor. We did not use an activation function for the Encoder output, and used sigmoidal activation for outputs from other networks. The channel size of the deconvolution block of the Discriminator was 129, not 128, because the usage of the minibatch standard deviation increased the channel size by one. Although we employed multiple Discriminators in our image deconstructor, they used the common architecture displayed here, but different parameters adjusted according to different loss functions. We did not use skip connections between the Encoder and the Decoder, and used deconvolution block, instead of full-connection block, for the output from the Encoder. These modifications were for making the latent space a three-dimensional tensor, and hence enabling the latent variable to convey information of local structures. Since there were two kinds of outputs from the Encoder, μ and σ^2, we employed two deconvolution blocks in parallel for the respective outputs. The channel size of the Decoder was 24, because it concatenated the middle-layer of the Map Generator (channel size $= 8$), and the output of the Encoder (dimensionality $= 16$); the latter was obtained by the reparametrization trick from its latent variable.

Appendix B. Transformed images obtained by BicycleGAN

Here, we examined if the most related existing image transformer, BicycleGAN, could be used for our objective. Fig. B.15 shows the transformed/generated images by the BicycleGAN, where left and right panels present the results when the original image was used and not used for training the Bicycle GAN, respectively. We can see that the reconstructed image was fairly apart from the original image, and in addition, the saliency map was not maintained in the image generated by the BycycleGAN from that of the original image. This was because the saliency map was of significant short of the information of the original image, and/or the latent dimensionality of the Bicycle GAN was as small as 8.

The similarity (in terms of Cosine Similarity (CS)) of the saliency map between the original image and the corresponding generated image was 0.883 ± 0.050 for 5000 validation images that were not used for training. This similarity was significantly smaller than that by our method (Table 2); in Fig. B.15, we actually see the saliency maps of some generated images were fairly different from those of their original images.

Fig. B.16 shows multiple images generated by probabilistic image generation with multiple random variables applied to the latent space of the BicycleGAN. The BicycleGAN tends to generate images with different color painting from that of the original image; this degenerated variability was due to the relatively low-dimensional latent space and the skip connections used in U-nets that consist of the whole image transformer.

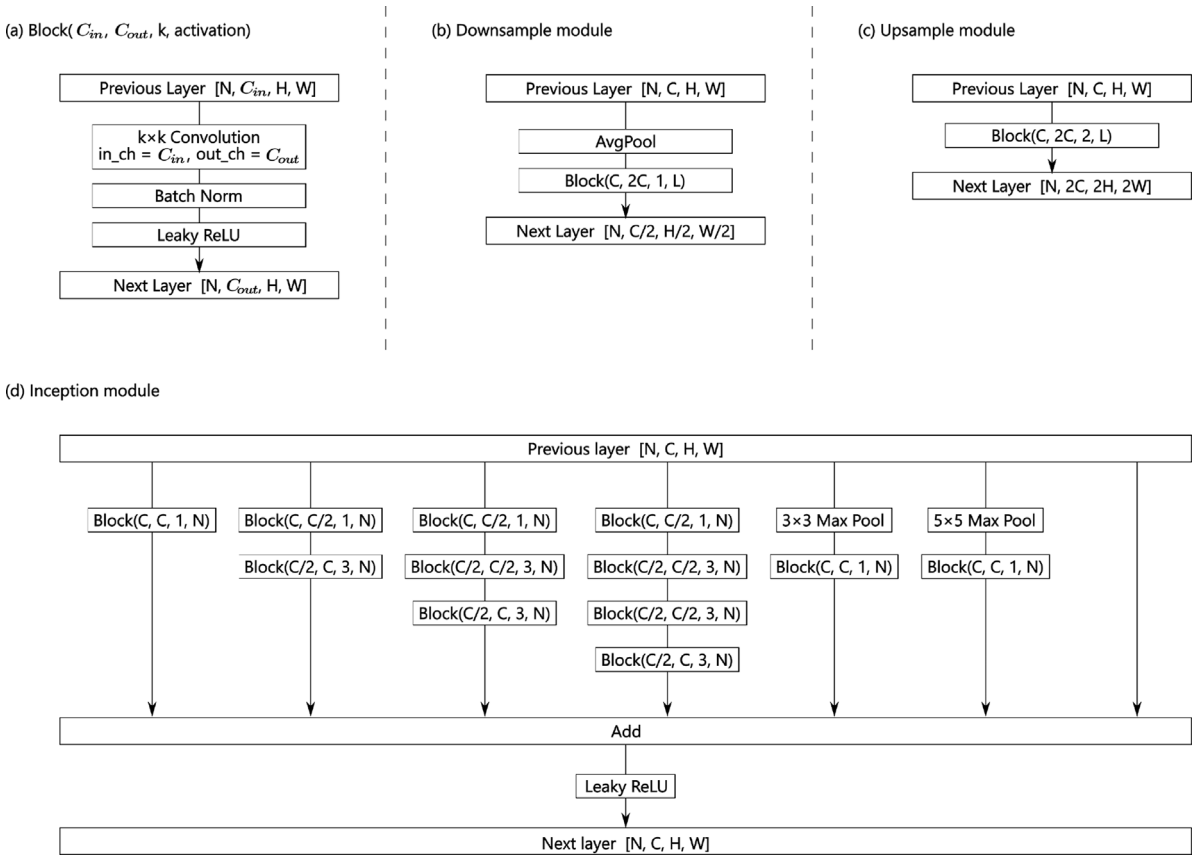

Fig. A.14. Specifications of modules used in our image deconstructor. (a) A Block is a single convolution layer, where C_{in}, C_{out}, and k denote the input channel size, the output channel size, and the kernel size, respectively. When an activation function is signified by L and N, we use a leaky ReLU activation and no activation (identity), respectively. (b) A downsample module. (c) An upsample module. (d) An inception module. $[N, C, H, W]$ next to a sign 'Previous layer' ('Next layer') means the size of input (output) features, where N is the sample number in the minibatch, and C, H, and W are the channel size, image height, and image width, respectively.

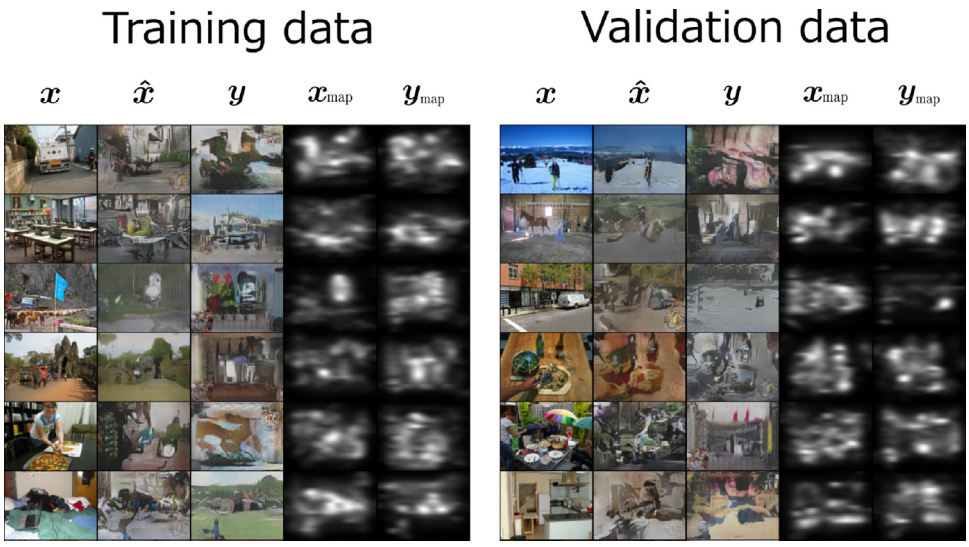

Fig. B.15. Image transformation by BicycleGAN. These results show when testing the BycycleGAN after being trained with the same image dataset as in our image deconstruction method, for several training images (a) and several validation images (b). The latter set of images was not used for training the BycycleGAN. Each panel shows an original natural image, a reconstructed image, a generated image, the saliency map of the original image, and the saliency map of the generated image, from the left to the right on each column.

Appendix C. Loss functions for the saliency map

Here, we examined how the generated images would depend on the loss function used for evaluating the saliency map. In our basic implementation, we used BCE, Eq. (10). Fig. C.17 shows the results when we used mean squared error (MSE) instead of the BCE in the training of the saliency map generator. Each column presents an original natural image, a reconstructed image,

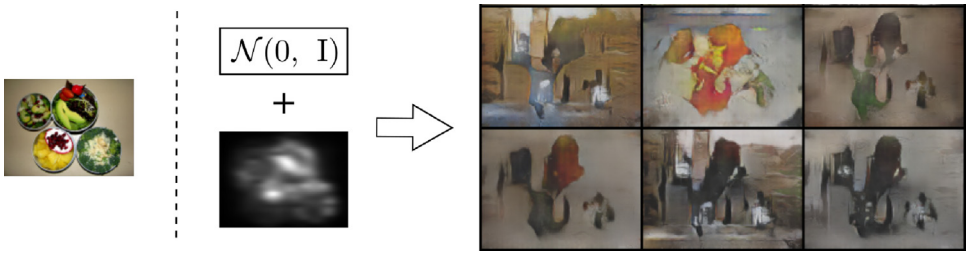

Fig. B.16. Diversity in the images transformed by the BicycleGAN. Due to the standard normal variable applied to the latent space, the BycycleGAN could produce images with a certain variability, but they were different mostly in the global features.

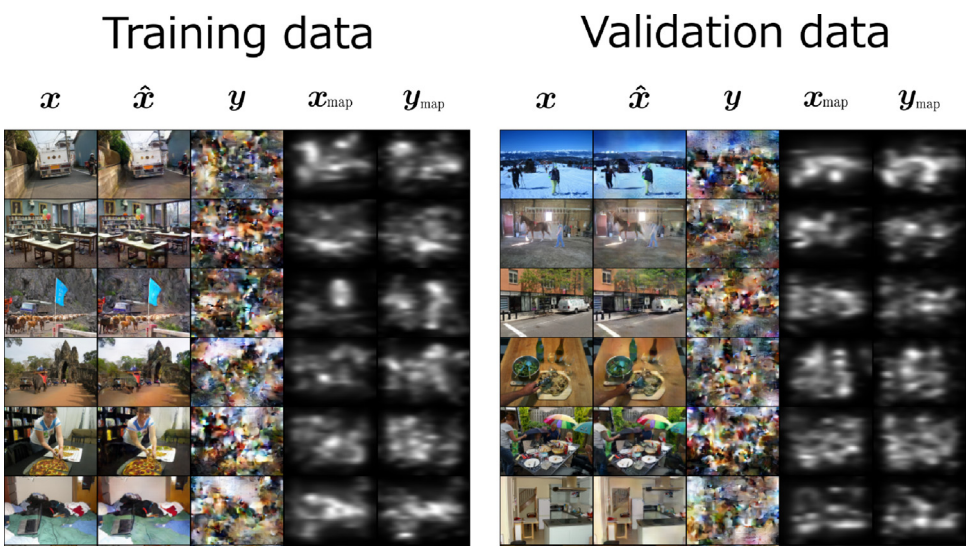

Fig. C.17. Image deconstruction when the loss function of the saliency map generator, Eq. (10) was replaced by mean squared error (MSE). We used the same visualization as in Fig. 4.

Table C.5
Cosine Similarity (CS) between saliency maps. This table shows the CS of the saliency maps, between of the original image and of the generated image. The comparison was done for two kinds of loss functions used for training the saliency map generator.

Model	CS
BCE	0.908 ± 0.029
MSE	0.898 ± 0.031

a generated image, the saliency map for the original image, and the saliency map for the generated image, from the left to the right on each row.

When we examined the similarity in terms of CS between the saliency map of an original image and that of the corresponding generated image, over the validation set of 5000 images, the training with the MSE was slightly lower than that with the BCE, but the difference was not significant (Table C.5 and Fig. C.17).

Appendix D. Further ablation study

Our image deconstructor used several techniques that had been proposed in the field of image transformation; i.e., feature matching, PatchGAN, minibatch standard deviation, and inception module, for improving the performance of our image deconstructor. Table 2 in the main text presents the results of a couple of ablation studies. Here, we present the results of further ablation studies; that is, each of the four techniques above was removed from the full model of our image deconstructor (Table D.6). This table suggests that the performance of the image deconstructor, in terms of the FID score, was substantially worse when we removed feature matching, PatchGAN, or inception module. We also found that minibatch standard deviation was minorly effective in our image deconstruction task.

Table D.6
Fréche Inception Distance (FID) (Heusel et al., 2017) in ablation studies. This table shows the FID score when either of the feature matching, PatchGAN, minibatch standard deviation, or inception module was removed from the full model of our image deconstructor.

Model	FID
Ours	322.95
Without feature matching	460.47
Without PatchGAN	415.27
Without minibatch standard deviation	321.01
Without inception module	382.15

Fig. D.18 shows the images generated by the four kinds of ablated models; the results are for validation images. Without the feature matching, the generated images were rough. Without the PatchGAN, the generated images were blurred, suggesting the PatchGAN was important for improving local structures of the generated images. Without the minibatch standard deviation, the generated images were a little distorted, which was commonly observed over many generated images. Although Table D.6 suggested a minor improvement by our full model over this ablated model, we consider the usage of the minibatch standard deviation was effective in removing such minor distortions. Without the inception module, the generated images became rough, suggesting the advantage of using the inception module in our image deconstruction method.

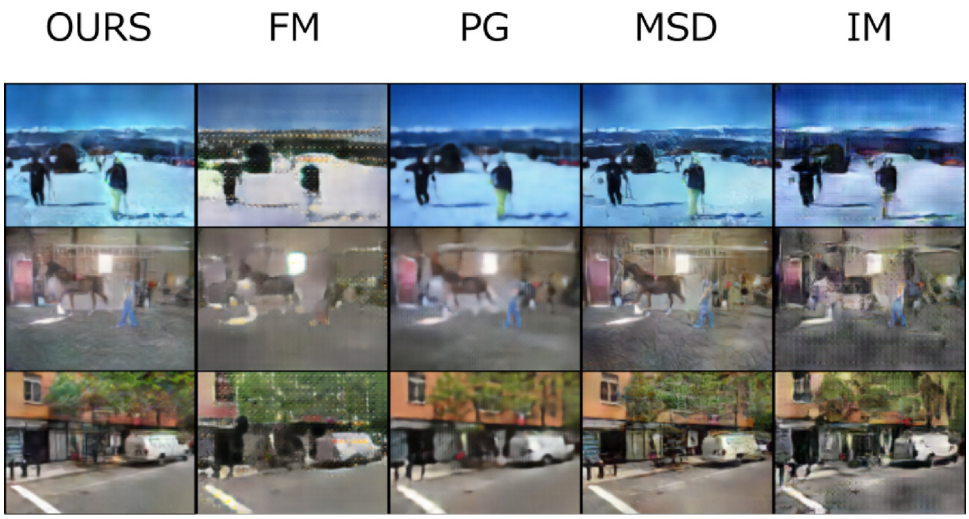

Fig. D.18. Images produced by our additional ablation studies. Images generated by the full model of our image deconstructor (OURS), and by ablated models without the feature matching (FM), PatchGAN (PG), minibatch standard deviation (MSD), or inception modules (IM), respectively. In the ablated model IM, INC(n) in Table A.4 was replaced with 'Block($n, n, 3, L$)'; that is, the inception module was replaced with a simpler convolution Block.

Appendix E. Implementations of Pix2Pix and AdaIN

E.1. Pix2Pix

Pix2Pix employs a typical GAN architecture; that is, there are a couple of modules, Generator and Discriminator, and the former is of an encoder–decoder architecture with the usage of U-net. The input and output of the Generator are a saliency map and its corresponding natural image, respectively. The natural image was taken from the SALICON dataset, and its corresponding saliency map was obtained by the Itti's method. This Generator was trained to lower a combined loss function of an L1 norm, which measures the discrepancy between the original natural image (for the input saliency map) and the image generated by the Generator, and an adversarial loss to fool the Discriminator. On the other hand, the Discriminator was trained to well discriminate between a pair of a natural (i.e., true) image and its saliency map and a pair of a generated (i.e., fake) image (output from the Generator) and its saliency map (input to the Generator). As well as in the original implementation of Pix2Pix, we also introduced Dropouts to the Generator training, which was found to be effective in attaining diversity in the generated images.

E.2. AdaIN

AdaIN has an encoder–decoder architecture, but there is no encoder training. The Encoder consisted of several convolution layers of VGG19 that was pretrained to perform well in the image classification task and then fixed. Although the architecture of the Decoder is of an upright one of the Encoder, the former was trained based on the loss function below. An input to Encoder was a pair of a content image and a style image, which were randomly chosen from a set of natural images taken from SALICON. After obtaining a couple of outputs, i.e., the feature vectors, of the Encoder, when input by a content image and a style image, we transformed the feature vector for the content image input such to make the feature-wise mean and variance to be consistent with those of the feature vector for the style image input. Then, this transformed feature vector was input to the Decoder. We used a loss function consisting of three terms, to train the Decoder; one is the content loss, which measures the Euclidian distance between the feature vector when the image generated by the Decoder was input to the Encoder and the feature vector input to the Decoder to generate the image; second is the style loss, the sum of the Euclidian distance of the mean and variance of activities of middle layer units between when the image generated by the Decoder and the style image were input to the Encoder; and, the maintenance loss of the saliency map, $\mathcal{L}_{\psi}^{\text{map}}$. Note that the third loss function was not used in the original AdaIN implementation.

References

Anzai, A., Ohzawa, I., & Freeman, R. D. (1999). Neural mechanisms for processing binocular information II. Complex cells. *Journal of Neurophysiology*, *82*(2), 909–924.

Bowman, S. R., Vilnis, L., Vinyals, O., Dai, A. M., Jozefowicz, R., & Bengio, S. (2015). Generating sentences from a continuous space. arXiv preprint arXiv:1511.06349.

Bylinskii, Z., Judd, T., Borji, A., Itti, L., Durand, F., Oliva, A., et al. MIT saliency benchmark, http://saliency.mit.edu/.

Dollar, P., Wojek, C., Schiele, B., & Perona, P. (2011). Pedestrian detection: An evaluation of the state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *34*(4), 743–761.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). Generative adversarial nets. In *Advances in neural information processing systems* (pp. 2672–2680).

Harel, J., Koch, C., & Perona, P. (2006). Graph-based visual saliency. *Advances in Neural Information Processing Systems*, *19*.

He, J., Spokoyny, D., Neubig, G., & Berg-Kirkpatrick, T. (2019). Lagging inference networks and posterior collapse in variational autoencoders. arXiv preprint arXiv:1901.05534.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in neural information processing systems* (pp. 6626–6637).

Huang, C.-W., Tan, S., Lacoste, A., & Courville, A. C. (2018). Improving explorability in variational inference with annealed variational objectives. In *Advances in neural information processing systems* (pp. 9701–9711).

Isola, P., Zhu, J.-Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1125–1134).

Itti, L. (2005). Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes. *Visual Cognition*, *12*(6), 1093–1123.

Itti, L. (2006). Quantitative modelling of perceptual salience at human eye position. *Visual Cognition*, *14*(4–8), 959–984.

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(11), 1254–1259.

Jiang, M., Huang, S., Duan, J., & Zhao, Q. (2015). Salicon: Saliency in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1072–1080).

Jiang, D., Li, G., Sun, Y., Hu, J., Yun, J., & Liu, Y. (2021). Manipulator grabbing position detection with information fusion of color image and depth image using deep learning. *Journal of Ambient Intelligence and Humanized Computing*, *12*(12), 10809–10822.

Karras, T., Aila, T., Laine, S., & Lehtinen, J. (2017). Progressive growing of gans for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114.

Kruthiventi, S. S., Ayush, K., & Babu, R. V. (2017). Deepfix: A fully convolutional neural network for predicting human eye fixations. *IEEE Transactions on Image Processing*, *26*(9), 4446–4456.

Larsen, A. B. L., Sønderby, S. K., Larochelle, H., & Winther, O. (2015). Autoencoding beyond pixels using a learned similarity metric. arXiv preprint arXiv:1512.09300.

Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I., & Frey, B. (2015). Adversarial autoencoders. arXiv preprint arXiv:1511.05644.

Mescheder, L., Nowozin, S., & Geiger, A. (2017). Adversarial variational bayes: Unifying variational autoencoders and generative adversarial networks. In *Proceedings of the 34th international conference on machine learning-Volume 70* (pp. 2391–2400). JMLR. org.

Ohzawa, I., & Freeman, R. D. (1986). The binocular organization of simple cells in the cat's visual cortex. *Journal of Neurophysiology*, *56*(1), 221–242.

Pan, J., Sayrol, E., Nieto, X. G.-i., Ferrer, C. C., Torres, J., McGuinness, K., et al. (2017). Salgan: Visual saliency prediction with adversarial networks. In *CVPR scene understanding workshop (SUNw)*.

Radford, A., Metz, L., & Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International conference on medical image computing and computer-assisted intervention* (pp. 234–241). Springer.

Rosca, M., Lakshminarayanan, B., Warde-Farley, D., & Mohamed, S. (2017). Variational approaches for auto-encoding generative adversarial networks. arXiv preprint arXiv:1706.04987.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., & Chen, X. (2016). Improved techniques for training gans. In *Advances in neural information processing systems* (pp. 2234–2242).

Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. A. (2017). Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first AAAI conference on artificial intelligence*.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2818–2826).

van den Oord, A., Vinyals, O., et al. (2017). Neural discrete representation learning. In *Advances in neural information processing systems* (pp. 6306–6315).

Veale, R., Hafed, Z. M., & Yoshida, M. (2017). How is visual salience computed in the brain? Insights from behaviour, neurobiology and modelling. *Philosophical Transactions of the Royal Society, Series B (Biological Sciences)*, *372*(1714), Article 20160113.

Weng, Y., Sun, Y., Jiang, D., Tao, B., Liu, Y., Yun, J., et al. (2021). Enhancement of real-time grasp detection by cascaded deep convolutional neural networks. *Concurrency Computations: Practice and Experience*, *33*(5), Article e5976.

Xu, J., & Durrett, G. (2018). Spherical latent spaces for stable variational autoencoders. arXiv preprint arXiv:1808.10805.

Yoshdia, M., Itti, L., Berg, D., Ikeda, T., Kato, R., Takaura, K., et al. (2010). Visually guided eye movements based on color saliency in monkeys with unilateral lesion of primary visual cortex. *Neuroscience Research*, (68), Article e100.

Zhu, J.-Y., Zhang, R., Pathak, D., Darrell, T., Efros, A. A., Wang, O., et al. (2017). Toward multimodal image-to-image translation. In *Advances in neural information processing systems* (pp. 465–476).